Handling Class Imbalance in Credit Card Fraud using Resampling Methods

Nur Farhana Hordri[1], Siti Sophiayati Yuhaniz[2],
Nurulhuda Firdaus Mohd Azmi[3]
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia
Kuala Lumpur, Malaysia

Siti Mariyam Shamsuddin[4]
UTM Big Data Centre
Universiti Teknologi Malaysia
Johor, Malaysia

***Abstract*—Credit card based online payments has grown intensely, compelling the financial organisations to implement and continuously improve their fraud detection system. However, credit card fraud dataset is heavily imbalanced and different types of misclassification errors may have different costs and it is essential to control them, to a certain degree, to compromise those errors. Classification techniques are the promising solutions to detect the fraud and non-fraud transactions. Unfortunately, in a certain condition, classification techniques do not perform well when it comes to huge numbers of differences in minority and majority cases. Hence in this study, resampling methods, Random Under Sampling, Random Over Sampling and Synthetic Minority Oversampling Technique, were applied in the credit card dataset to overcome the rare events in the dataset. Then, the three resampled datasets were classified using classification techniques. The performances were measured by their sensitivity, specificity, accuracy, precision, area under curve (AUC) and error rate. The findings disclosed that by resampling the dataset, the models were more practicable, gave better performance and were statistically better.**

***Keywords*—Credit card; imbalanced dataset; misclassification error; resampling methods; random undersampling; random oversampling; synthetic minority oversampling technique**

I. Introduction

In the past decades when businesses were migrated and evolved to the online business and money was managed electronically in an ever-growing cashless banking economy, credit cards were gradually replacing the use of cash over its suitability [1]. Credit cards became the most popular mode of payment ever since. According to [2], credit card-based purchases can be categorised into two types: i) physical card purchase and ii) virtual card purchase. Most payments for online purchases were under virtual card purchases which few information were needed such as card numbers, expiration data, and secure codes. Along with the increasing numbers of the credit card users, the numbers of fraudulent transactions have been constantly increased. In the article [3] stated that, it is hard to find the identity and the location of the fraudsters since the evidences were hidden behind the internet. The merchants that were facing with the credit card fraudsters will bear all the costs including card issuer fees, charges, and administrative charges [4]. Consequently, the merchants must increase the price of the goods or give more discounts or reduce the incentives to conceal all the losses. Hence, an effective fraud detection system is vital to reduce the losses rate.

Before proceeding with the fraud detection system, it must be bear in mind that there had been an enormous increase in the amount of credit card dataset collected and processed by the organisations. Normally in the real dataset, the number of fraudulent is very rare as compared with the non-fraudulent transactions [5, 6, 7]. Conceivably with a skewed dataset, the performance of the system surely drops in terms of its accuracy. When a legitimate transaction is misclassified as a fraudulent transaction, it will affect the customer services and causes to lose trust from the customers and the financial institution [8, 9, 10]. Maes (2002) have provided some capacity that a fraud detection system should have in order to perform a good result [11]. The system should be able to: i) handle skewed distributions, ii) handle noise, iii) avoid the overlapping data iv) adapt themselves to new kinds of frauds, v) evaluate the classifier using good metrics, and vi) detect the behaviour of the frauds. Recent research in [12] stated there are three challenges to construct the fraud detection system. The challenges are: i) the data distribution evolves over time because of seasonality and new attack strategies ii) fraudulent transactions represent only a very small fraction of all the daily transactions and iii) the fraud detection problem is intrinsically a sequential classification task.

In 2017, Haixian et al. stated that it is difficult to detect the rare events due to their infrequency and casualness. Plus, it can result in heavy cost if misclassified the rare events. In their review paper, they have identified three main solutions to the challenges: resampling, cost-sensitive learning, and ensemble methods [13]. The most popular method is resampling methods which are used to rebalance the imbalanced dataset in order to alleviate the effect of the skewed class distribution in the learning process. Secondly, cost-sensitive learning which can be incorporated to both data level and algorithmic level. Lastly, ensemble method is used to improve the performance of a single classifiers that outperform. In a review paper [14], they specified two approaches should be performed to solve the imbalanced data problems: i) solution at data level by balancing the distribution of the majority and minority class trough methods of under sampling, over sampling or combination of both methods, ii) solution at algorithm level by modification in classifier methods or optimise the performance of learning algorithm.

Thus in [13 and 14], both review papers emphasized that there are no absolute methods that are more efficient in dealing with the class imbalance. They found out some insights about commonly-used methods in some domains.

Burez (2009) handled the imbalanced class in churn prediction by applying several methods. The methods are: i) evaluation metrics, ii) cost-sensitive learning, ii) resampling methods and iv) boosting. He used ROC analysis as for evaluation metrics and stochastic gradient boosting learner as for boosting. For cost-sensitive learning, he used random forest and random under sampling as for resampling methods [15]. The study in [16] proposed an efficient resampling method and obtained comparable classification results between random under sampling and random over sampling. The experiments were carried out using four large imbalanced Bioinformatics datasets. They have recommended 100%-under(0.75)-over method for obtaining comparable classification results to the over sampling results. In 2002, [17] has proposed Wrapper-based Random Oversampling (WRO) to handle class imbalanced problem. Wrapper is a pre-processing method that incorporates the classifier output to guide pre-processing. The method oversampled the minority class data randomly and the classifier is optimised. They evaluated the WRO with real dataset that they obtained from UCI repository. WRO has better results in most experiment compared to Synthetic Minority Over Sampling Techniques (SMOTE) and random over sampling. Research in [18] investigated the resampling methods specifically on data from Spotify users. They used the most common oversampling methods: random oversampling and SMOTE, and the most common under sampling method: random under sampling. Yan and Han (2018) proposed RE-sample and Cost-Sensitive Stacked Generalisation (RECSG) based on 2-layer learning models to solve the imbalanced problem in 18 benchmark datasets [19]. The experimental results and statistical tests showed that the RECSG approach improved the classification performance.

In reviewing the literature, resampling methods is the main focus of this study due to its simplicity and compatibility with existing classification models to handle the rarity event in massive credit card dataset. There is no research yet were found on the association between credit card fraud and resampling methods. Therefore, the aim of this study is to investigate the classification models' ability to classify the fraud and non-fraud transactions, and to examine if the different resampling methods could improve the performances of the models. The research methodology of the study is conducted in Section 2. Thereafter in Section 3, the experimental setup is described. Next, the results and discussions is presented in Section 4. This study ends with conclusion remarks and future works in Section 5.

II. Research Methodology

This section gives brief description of the methodology of this study. In addition, this section also discusses each step of the methodology. Fig. 1 displays the framework of research methodology of this study.

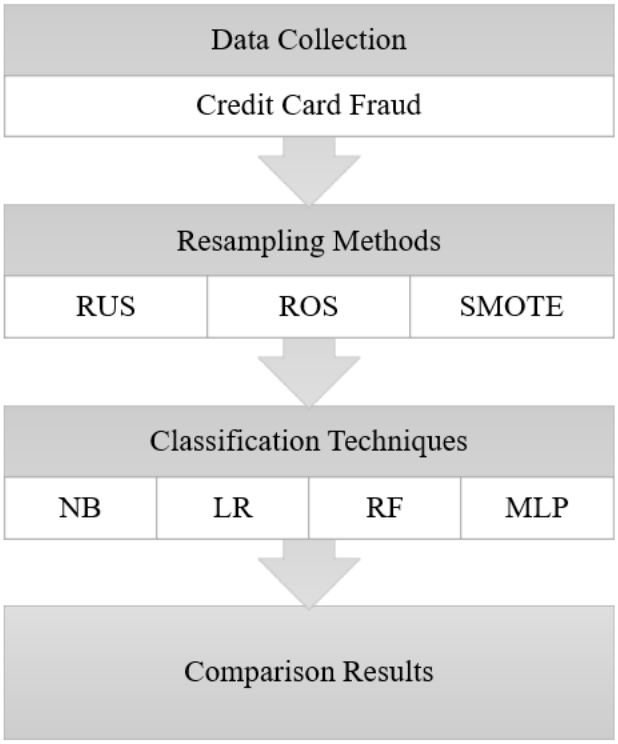

Fig. 1. Framework of Research Methodology.

A. Data Collection

One of the biggest problems associated with researchers in financial fraud detection is lack of real-life data because of sensitivity of data and privacy issue [5]. Hence, a publicly available dataset is downloaded from [20] to be used in this research. It has a total of 284,807 transactions made in September 2013 by European cardholders. The dataset contains 492 fraud transactions, which is highly imbalance.

B. Resampling Methods

Three widely-used methods for resampling in this study are Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE. For undersampling, RUS is chosen, since it is considered both simple yet effective. ROS and SMOTE were chosen as oversampling methods because of its widely usage. Furthermore, ROS is an intuitive way of balancing data, whereas SMOTE is more complex creating synthetic samples using K-Nearest Neighbour (KNN). Table 1 below summarises the differences between the three resampling methods.

C. Classification Techniques

Credit Card dataset is a binary classification task. Either the transaction is classified as non-fraud (0) or fraud (1). After the data have been resampled accordingly, the models are needed to be trained using classifiers to evaluate the methods. Thus, in this study, four different classification techniques were explored: Naïve Bayes (NB), Linear Regression (LR), Random Forest (RF) and Multilayer Perceptron (MLP). A summary of the strength and limitations of the classifiers used in this study is given in Table II.

TABLE I. COMPARISON OF RESAMPLING METHODS

	RUS	ROS	SMOTE
Process	Shrink the majority data by randomly discarded the data from the dataset.	Expand the minority data by duplicated the data randomly.	Expand the minority data by extrapolating between preexisting minority instances which obtained by KNN.
Strength	Shorter convergence time.	Not possess any information loss, could produce better results.	Effective in improving the classification accuracy of the minority data.
Limitation	Loss the important information.	Trapped with overfitting due to multiple tied instances.	Data synthetic still possible to spread on both minority and majority data, hence reduced the performance of classification.

TABLE II. STRENGTH AND LIMITATION OF THE CLASSIFICATION TECHNIQUES

Classifier	Strength	Limitation
Naïve Bayes	Simplest classifiers and perform well in classification.	Assumes that all attributes are independent of each other given the context of the class.
Linear Regression	Provide optimal results when the relationship between independent and dependent variable are almost linear.	Sensitive to outliers and limited to numeric values only.
Random Forest	Require low computational power and suitable for real-time operations. The procedure is easy to understand and implement.	Easily to get overfitting when training set does not give underlying domain information. Whenever have new types of cases need retrain.
Multilayer Perceptron	Suitable for binary classification problems.	Retraining is required and need high computational power. Unsuitable for real-time operations.

III. EXPERIMENTAL SETUP

This section describes the division of the data in training dataset and the performance measures conducted throughout of this study. All the resampling techniques are implemented in Java framework of WEKA 3.8 for comparative evaluations. The parameters for the classification techniques were set accordingly by default. No further fine tuning of parameters to specific datasets can be beneficial, consideration of generally accepted settings is more typical in practice.

A. Data Division

The methodological approach taken by this study is motivated from research [15]. The research handled the imbalance problem in churn prediction by resampling the minority and majority classes based on ratio 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20 and 90:10 where churners proportionate with non-churners. Due to the results that the research have obtained and limitation of time, this study chose ratio 30:70 and ratio 50:50 (fraud:non-fraud) to divide the training dataset for this study. An overview of the dataset division, splitting and resampling, can be seen in Fig. 2. Following Fig. 2 is Table III and IV which have more details on dataset division for this research.

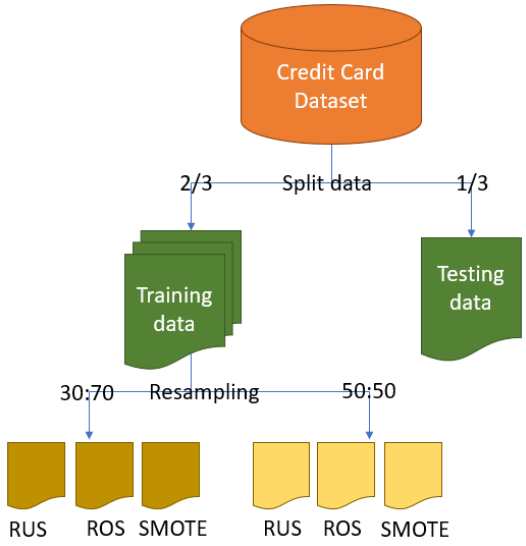

Fig. 2. Division of Dataset.

TABLE III. DIVISION DATASET BY RATIO 30:70

Data Division	Training Dataset	Resampling Method		
		RUS	*ROS*	*SMOTE*
Fraud	328	328	81233	81299
Non-Fraud	189544	768	189544	189543
Total	189872	1097	270777	270842

TABLE IV. DIVISION DATASET BY RATIO 50:50

Data Division	Training Dataset	Resampling Method		
		RUS	*ROS*	*SMOTE*
Fraud	328	328	189544	189543
Non-Fraud	189544	328	189544	189543
Total	189872	656	379088	379086

B. Performances Evaluation

In this study, performance evaluations were conducted to assess the performance of the classification methods for each resampling technique. The models have two fundamental errors may occur: classifying a fraud falsely as a non-fraud and classifying a non-fraud falsely as a fraud. These errors are more commonly known as false positive and false negative results. Other possible classifications will be correctly classified such as true positive and true negative results. The correlation between these are presented in a confusion matrix in Table V. Performance of four classifiers were compared in terms of Sensitivity, Specificity, Accuracy, F-Measure and Area Under Curve (AUC). These metrics are calculated using the confusion matrix as shown below.

TABLE V. CONFUSION MATRIX OF CREDIT CARD DATASET

	Classified as Fraud	**Classified as Non-Fraud**
Fraud	True Positive (TP)	False Negative (FN)
Non-Fraud	False Positive (FP)	True Negative (TN)

Table V was generated from the four measures: True Positive (TP) — the number of correctly classified as a fraud and it is really a fraud, True Negative (TN) — the number of correctly classified as non-fraud and it is really a non-fraud, False Positive (FP) — instances which were incorrectly classified as a fraud but it is a non-fraud and False Negative (FN) — instances which were incorrectly classified as non-fraud but it is a fraud.

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{\sum(TP+FP+TN+FN)} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

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

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

$$Error = \frac{FP+FN}{\sum(TP+FP+TN+FN)} \quad (7)$$

IV. RESULTS AND DISCUSSIONS

This section discusses the results that were obtained from the experiments. Table VI and Table VII displays the summary of the comparison results for each classification techniques in three resampling methods by ratio 30:70 and 50:50, correspondingly. The results were compared in terms of sensitivity, specificity, accuracy, precision, F-measure, AUC and time taken to build the model in seconds. All the classifiers were performed well with an accuracy of 0.90 or more. Though, RF dominates with higher accuracy compared to other classification techniques for both ratios of each resampling method.

Table VI provides the information of ratio 30:70 for three resampling techniques. As can be seen in RUS, MLP has higher sensitivity if compared to other classification techniques but have slightly lower specificity than RF. The error rate for MLP and RF are 0.0319 and 0.0273, correspondingly. Thus, RF have approximately 2% of misclassification rate compare to MLP which is have 3% of misclassification rate. For ROS, LR and RF have accuracy 99% which they can correctly identified the fraud and non-fraud of the credit card dataset. However, RF is more precise compared to LR. On the other hand, LR only took 53.5 seconds to build the model while RF took about 343 seconds. The longest time taken to build the model is MLP which is 896 seconds. Meanwhile in SMOTE, RF has higher precision rate compare to other classification methods. It shows that RF often correctly classified non-fraud dataset with 0.9999 rate. Followed by LR (0.9862), MLP (0.9837) and NB (0.9328).

TABLE VI. COMPARISON RESULTS OF CLASSIFICATION TECHNIQUES BY RATIO 30:70

Resampling Methods: Random Under Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.85321101	0.97009103	0.93521898	0.92384106	0.88712242	0
Linear Regression	0.89602446	0.9869961	0.95985401	0.9669967	0.93015873	0.08
Random Forest	0.90825688	1	**0.97262774**	1	0.95192308	0.5
Multilayer Perceptron	0.9204893	0.98829649	0.96806569	0.97096774	0.94505495	4.52
Resampling Methods: Random Over Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.84758845	0.97382652	0.93595468	0.93279053	0.88815077	2.11
Linear Regression	1	0.99763642	0.9983455	0.99451532	0.99725012	53.5
Random Forest	1	1	**0.99995199**	1	1	343.43
Multilayer Perceptron	0.96569171	0.9989079	0.98894293	0.99736822	0.98127439	896.31
Resampling Methods: Synthetic Minority Over Sampling Technique						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.833294	0.97426441	0.93194925	0.93283212	0.88025831	2.51
Linear Regression	0.98990147	0.9940594	0.99281131	0.98620166	0.9880481	88.16
Random Forest	0.99924968	0.99957266	**0.99947571**	0.99900392	0.99912679	716.2
Multilayer Perceptron	0.98723232	0.99297785	0.9912532	0.98368713	0.98545653	1034.57

TABLE VII. COMPARISON RESULTS OF CLASSIFICATION TECHNIQUES BY RATIO 50:50

Resampling Methods: Random Undersampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.85321101	0.95731707	0.90534351	0.95221843	0.9	0
Linear Regression	0.91131498	0.97865854	0.94503817	0.97704918	0.94303797	0.05
Random Forest	0.92966361	0.97865854	**0.95419847**	0.97749196	0.95297806	0.45
Multilayer Perceptron	0.9204893	0.96036585	0.94045802	0.95859873	0.93915757	2.63
Resampling Methods: Random Oversampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.8497982	0.97318287	0.91149021	0.96940864	0.9056713	2.91
Linear Regression	1	0.99393278	0.9969664	0.99396943	0.99697559	72.17
Random Forest	1	0.99992614	**0.99996307**	0.99992614	0.99996307	451.36
Multilayer Perceptron	0.98128149	0.99772611	0.98950376	0.99768812	0.98941679	1536.34
Resampling Methods: Synthetic Minority Oversampling Technique						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.8335048	0.97364186	0.90357333	0.96934612	0.89630777	4.5
Linear Regression	0.99559467	0.99172747	0.99366107	0.99175934	0.9936733	131.6
Random Forest	0.99996307	0.99952517	**0.99974412**	0.99952538	0.99974418	991.69
Multilayer Perceptron	0.99217592	0.99018692	0.99118142	0.9902064	0.99119018	1499.7

In contrast, Table VII presents the information of ratio 50:50 for each resampling method. In RUS, LR and RF have similar specificity rate which is 0.97866 but different in sensitivity rate nearly at 0.0183. Although both techniques can classify the same number of fraud dataset, RF is still better in classifying the fraud dataset compare to LR. RF also expresses the effectiveness of classification in terms of high F-Measure compared to other classifiers. In the meantime, for ROS, RF has higher accuracy rate. Follows by LR, MLP and NB with 0.9969, 0.9895 and 0.9115, subsequently. Both LR and RF have equal numbers of sensitivity rate which is 1. This means that LR and RF have 100% correctly classified the fraud dataset. Meanwhile, RF gives comprehensive results even though in SMOTE. RF has a small differences of accuracy rate compared to LR and MLP with 0.00608 and 0.00856. Hence, RF can correctly classified fraud and non-fraud in the dataset since it has higher sensitivity rate and precision rate compared to other classification methods. It is important to view that, although NB only took a split of second to build the model, it has the lowest precision rate compared to the rest. NB also has the lowest accuracy rate and sensitivity rate. Albeit MLP has the longest time taken to build the model, MLP is doing well in classifying the fraud and non-fraud which add up to 99% correctness.

Table VIII is quite revealing in several ways. First, unlike the other tables, Table VIII is more focusing on performance of AUC and error rate. Secondly, RUS, ROS and SMOTE were compared with the original training dataset. The highest AUC and lowest error were printed in bold. For each of the case, the test statistics with the highest AUC and lowest errors were calculated and compared with models that were significantly worse. It can be seen in the table that although the original training set should have the lowest rate compared to other resampling methods in four classifiers, it has the worst performance in AUC. It is an example that the model does not have a good statistic.

For NB, the best AUC is obtained by the RUS when the training set is set to 30% of fraud and 70% of non-fraud (AUC=0.9117). It is only has 0.0002 differences compared to ROS when the training sets have the same ratio of fraud and non-fraud. SMOTE and the original training set do not differ significantly what concerns to AUC. While for LR, the closest AUC to 1 is ROS by ratio 30:70 with 0.9988. Followed by the ROS (50:50), SMOTE (50:50) and SMOTE (30:70). RUS in both ratios and the original training set have large differences in AUC compared to ROS with 30% fraud. Although the original training sets have the lowest error set, ROS (30:70) has the second lowest error rate. ROS with 30% of fraud and 70% non-fraud is significantly better in statistic, therefore it can be a better resampling method for LR technique.

TABLE VIII. MEAN AUC PERFORMANCE AND ERROR RATE FOR EACH RESAMPLING METHODS

Techniques: Naïve Bayes							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.89271531	**0.91165102**	0.90526404	0.91070748	0.91149053	0.9037794	0.90357333
Error	**0.02278916**	0.06478102	0.09465649	**0.06404532**	0.0885098	0.06805075	0.09642667
Techniques: Linear Regression							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.87036088	0.94151028	0.94498676	**0.99881821**	0.99696639	0.99198044	0.99366107
Error	**0.00057934**	0.04014599	0.05496183	**0.00165450**	0.0030336	0.007188693	0.006338931
Techniques: Random Forest							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.88259557	0.95412844	0.95416107	**1**	0.99996307	0.99941117	0.99974412
Error	**0.00045821**	0.02737222	0.04580153	0.00004801	**0.00003693**	0.000524291	0.00025588
Techniques: Multilayer Perceptron							
	Training Data	**RUS**		**ROS**		**SMOTE**	
Ratio		30:70	50:50	30:70	50:50	30:70	50:50
AUC	0.88258238	0.95439289	0.94042758	0.9822998	0.9895038	0.99010508	**0.99118142**
Error	**0.00048454**	0.03193431	0.05954198	0.01105707	0.01049624	**0.00874680**	0.00881858

Similar in LR, the better performance of AUC using RF classifier is ROS by ratio 30:70 with 1. ROS (50:50) is following closely with AUC = 0.99996. From the table, ROS (50:50) have the smallest error rate as well after the original training set. Both ratios in ROS show significantly better in statistic for RF. When looking at MLP, what concerns of the performance of AUC, SMOTE gave better result in both ratios. Followed by ROS and RUS. Similar to error rate where SMOTE also gave the smallest rate compared to ROS and RUS. Yet, none of the resampling methods were significantly better in terms of error rate than the original training set.

V. CONCLUSIONS AND FUTURE WORK

This study was set out with the aim to investigate the classification models' ability to classify the fraud and non-fraud transactions, and to examine if the different resampling methods could improve the performances of the models. It is interesting to note that in all four classifiers that have been applied, RF showed a robust performance in three resampling methods. RF succeeded to get higher accuracy compared to NB, LR and MLP for the resampling methods. It would be interesting to compare the classification techniques used in this study with other techniques such as Support Vector Machines, Neural Network and Genetic Algorithm.

Surprisingly, it has been found out that ROS was found to give convincing results if compared to SMOTE. Although SMOTE is quite effective in the literature, this is most probably due to some of the synthetic data resulting from the oversampling process were spreading on both minority and majority data, as discussed in Section 2, which is the main limitation of SMOTE. There were few researchers that have modified SMOTE to create more effective methods in improving the classification performance. Perhaps, this improve-SMOTE can be compared with the current resampling methods for credit card dataset in the future. Hence, these results may provide further support to the organisation to build better fraud detection system which can handle the skewed distribution and noise as well to evaluate the classifier using better metrics.

ACKNOWLEDGMENT

The authors honorably appreciate the Ministry of Education Malaysia, Universiti Teknologi Malaysia (UTM), UTM Big Data Centre and MyBrain15 for their support. This work is funded by Fundamental Research Grant Scheme (FRGS) with the title "Enhancing Data Analytics Algorithms using Deep Learning Approaches in Predicting Big Data Cyber-Enabled Crimes" (4F877).

REFERENCES

[1] Ki, Y., & Yoon, J. W. (2018, January). PD-FDS: Purchase Density based Online Credit Card Fraud Detection System. In KDD 2017 Workshop on Anomaly Detection in Finance (pp. 76-84).

[2] Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), 937-953.

[3] Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE ACCESS, 6, 14277-14284.

[4] Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert systems with applications, 35(4), 1721-1732.

[5] Reddy, M. S., Indraja, S., & Nikhil, L. (2017). Implementation Of Neural Network For Cashless Transactions In Credit Card Transactions. International Journal of Recent Trends in Engineering & Research.

[6] Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015, September). When is undersampling effective in unbalanced classification tasks?. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 200-215). Springer, Cham.

[7] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In Computational Intelligence, 2015 IEEE Symposium Series on (pp. 159-166). IEEE.

[8] Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and Their Applications, 14(6), 67-74.

[9] Chan, P. K., & Stolfo, S. J. (1998, August). Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. In KDD (Vol. 98, pp. 164-168).

[10] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. Intelligent data analysis, 6(5), 429-449.

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

[12] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems, 29(8), 3784-3797.

[13] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

[14] Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017, March). Synthetic Over Sampling Methods for Handling Class Imbalanced Problems: A Review. In IOP Conference Series: Earth and Environmental Science (Vol. 58, No. 1, p. 012031). IOP Publishing.

[15] Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

[16] Batuwita, R., & Palade, V. (2010, July). Efficient resampling methods for training support vector machines with imbalanced datasets. In Neural Networks (IJCNN), The 2010 International Joint Conference on (pp. 1-8). IEEE.

[17] Ghazikhani, A., Yazdi, H. S., & Monsefi, R. (2012, May). Class imbalance handling using wrapper-based random oversampling. In Electrical Engineering (ICEE), 2012 20th Iranian Conference on (pp. 611-616). IEEE.

[18] Jagelid, M., & Movin, M. (2017). A Comparison of Resampling Techniques to Handle the Class Imbalance Problem in Machine Learning: Conversion prediction of Spotify Users-A Case Study.

[19] Yan, J., & Han, S. (2018). Classifying Imbalanced Data Sets by a Novel RE-Sample and Cost-Sensitive Stacked Generalization Method. Mathematical Problems in Engineering, 2018.

[20] Credit Card Fraud Detection. Accessed: June. 7, 2018. [Online]. Available: https://www.kaggle.com/dalpozz/creditcardfraud