Oversampling Techniques in Machine Learning Detection of Credit Card Fraud

Charlie Obimbo, Davleen Mand, Simarjeet Singh
University of Guelph, Canada

Abstract

More than ever before, the trend of doing things online has been explored and successfully implemented in many areas, including online shopping, online learning, working online, to name but a few. However, it has brought with it challenges, including the fraudulent use of credit cards in online purchases, the challenge of academic integrity in online learning, especially in doing exams online, and how to keep people in engaged in meetings, when working and studying online, and still give them adequate privacy. This paper deals with the attempt to detect the fraudulent use of credit cards in a timely manner, to avoid as much negative effects in the world of E-commerce and help maintain consumer confidence. Thus, in the current study, machine learning algorithm LightGBM has been used to detect fraudulent credit card transactions from a real-life dataset containing credit card transactions of the customers. The performance of this classifier is compared with two state-of-the-art classifiers – Decision Tree, and Random Forests, which are extensively used for solving such problems. Since there is data imbalance between fraudulent and nonfraudulent class, the data sampling technique used is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE Oversampling performed best on all classifiers and LightGBM obtained precision value of 1 for both fraudulent and non-fraudulent class.

1. Introduction

In recent years, E-commerce has become an essential part of the global retail landscape. Like many other industries, with the advent of the Internet, continued digitization of modern life, and the attempt of humanity to keep safe with the emergence of the current deadly pandemic of Covid 19 that has been ravaging the world for now almost two years, consumers in virtually every country are now reaping the benefits of online transactions. As internet access and adoption increase rapidly around the world, the number of digital shoppers continues to increase every year. In 2020, more than two billion people purchased goods or services online, and in the same year, retail sales exceeded US $4.2 trillion globally. Figure 1 below shows the retail E-commerce worldwide sales between 2014 and 2021, and also the projection for 2022, according to Statista [1]. However, with this has also increased the cases of credit card fraud. As can be seen in Figure 2 below [2].

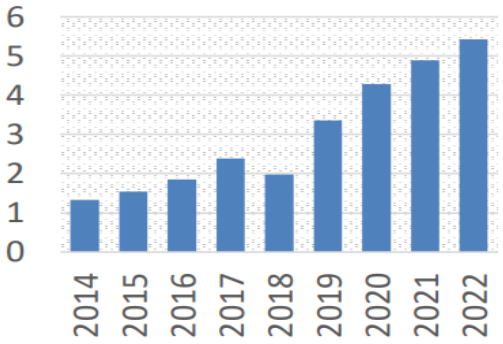


Figure 1. Retail E-commerce sales from 2014-22 (in Trillions (109) of US$)

Figure 2. Trends in credit card fraud from Q4'20 to Q1'21 [2]

Some of the Credit card frauds may be categorized as follows [3]:

1. Application Frauds: A fraudster may get control of an application system by gaining access to critical user information such as passwords and usernames and creating a false account. It frequently occurs in the

 | DOI: 10.20533/jitst.2046.3723.2021.0091

context of identity theft. The fraudster then requests for credit or a new credit card in the cardholder's name. In order to support or substantiate their bogus application, the fraudster steals the supporting documents.

2. CNP (Card Not Present): This may occur, when the perpetrator knows sensitive information of the card, such as the card number, expiry date and CVV without actually having the card itself.

3. Counterfeit Card Fraud: Skimming is a common method for making a Counterfeit Card. A false magnetic swipe card is created, which has all of the information from the genuine card. The forged card is completely functioning and can be used to conduct future purchases.

4. Lost and Stolen Card Fraud: When the original cardholder misplaces their card, it can fall into the hands of fraudsters, who can then use it to make purchases. It is difficult to do this using a machine because of the necessity of a pin, but internet transactions are simple enough for a fraudster. Card information may also be obtained using scanners (which maybe either home-made or purchased). These may be used to, obtain the cardholder's name and card number without contact. Say, from a distance of 10 feet, a cell phone-sized RFID reader powered at 30 dBm (decibels per milli-watt) would be able to do this.

5. Account Takeover: This is one of the most common types of deception. The fraudster has access to 402's account information. The actual card holder, as well as several important documents, are published by Blue Eyes Intelligence Engineering and Sciences Publication. The credit card company is then called up, and the fraudster emulates the original cardholder, and even requests an address change. Proof may even be presented by information obtained through social engineering. The duplicate card is then mailed to the new or fictitious address, which the offender can use.

6. False Merchant Sites: in this case, like the phishing attack, the customer is lured to fake webpage, created by a fraudster. This site resembles the genuine one. Discounts may be offered as bait, and when the customer attempts to purchase, all the customers information is gathered, and this may be used to perform fraudulent exchanges.

7. Merchant Collusion: Merchants may be in cahoots with fraudsters, to deliberately pass on credit card information.

The motivation for this work is to develop a classifier that will help to detect fraudulent transactions, which in turn will help in reduction of fraudulent activities in future as these will be captured early which will prevent customers from paying for unwanted transactions. This will also help to save millions of dollars lost by financial institutions due to such illegal activities. The biggest challenge in current study is large data imbalance between non-fraudulent class (majority) and fraudulent class (minority) as it leads to ignorance of minority class [5] and classifier keeps predicting majority class [6]. To tackle this challenge, three data sampling techniques are used; random oversampling, random under sampling and SMOTE oversampling. Random oversampling technique removes the class imbalance by replicating random examples of minority class [7]. But this sampling might lead to overfitting of the model as it learns from same data samples multiple times. In random under sampling, random data points of majority class are removed from dataset. But removing datapoints leads to loss of information [8] and underfitting of the model on training dataset. In SMOTE oversampling, data points of minority class are oversampled by “introducing synthetic examples along the line segments joining any of the k minority class nearest neighbors” [9]. SMOTE helps in preventing overfitting of the model and provides better generalization as it spreads the decision boundary between majority and minority class [10].

LightGBM is a gradient boosting framework which uses histogram-based algorithms and has better performance [11]. Gradient Boosting is an ensemble of decision trees; it combines several shallow trees, which are weak learners into a strong classifier. On the other hand, decision trees perform extremely well in binary classification problems. The approach used in decision trees is breaking down complex decision into a series of simple decisions to arrive at desired solution [12]. In Random Forests classifier, multiple trees are created, and each tree is trained on bootstrapped sample of training data [13]. For predicting the output, each tree casts a vote and class with maximum number of votes wins and is predicted as output label by the classifier [13]. F-score and precision values give a clearer insight about performance of a classifier on imbalanced datasets [14]. Since the dataset in current study is heavily biased towards fraudulent class, F-score and precision values are considered while evaluating performance of three classifiers (Decision Tress, Random Forests, LightGBM) rather than accuracy value.

 | DOI: 10.20533/jitst.2046.3723.2021.0091

2. Dataset Used

The dataset is obtained from Kaggle. It has transactions made by European cardholders in September 2013 [15]. Out of 284,807 transactions in the dataset, 492 are fraud which works out to be 0.172% of all transactions. In order to provide privacy of customers, all columns except transaction amount and time between transactions are transformed using Principal Component Analysis (PCA). Class of fraud transaction is labelled as 1 and genuine transactions is labelled as 0.

3. Background and Methodology

In this section, a review of literature related to credit card fraud detection is done.

3.1. Literature Review

Several researchers have used super-apps to enhance credit card fraud detection. Super apps are digital platforms that contain many different services and can analyze user-profiles. Some of the services provided by Super App include, but are not limited to, financial services, food delivery services, markets, and travel services. The super application uses the collected interactions to enhance the performance of the model in multiple domains [16], [17], [18], [19]. This model has gained great traction in the market. Companies such as WeChat and Alipay have achieved gratifying results in analyzing financial behaviour patterns without traditional financial data.[3].

Hui Han et. al. [20] used SMOTE and borderline SMOTE oversampling techniques to remove the class imbalance on four datasets: Circle, Pima, Satimage, and Haberman. F-value and True Positive (TP) rate were used as performance metrics. Borderline SMOTE just oversamples minority data points which are along the decision boundary. Experimental results showed borderline SMOTE outperformed traditional SMOTE as former was able to achieve higher F-value and TP value on all four datasets.

3.2. Methodology

In supervised learning, the classifier is first trained on training dataset and its accuracy is then tested on test dataset. The current study deals with binary classification problem and follows supervised learning approach. The spilt ratio chosen for training-test dataset is 70-30. Choosing a higher split value of either 80-20 or 90-10 might lead to overfitting of the model as number of datapoints is quite high. A model which is overfit on training data leads to poor generalization [19]. It is important to normalize all features in a dataset and make them in uniform size [20]. Hence all features in current dataset are normalized. Training a model with normalized features is less computationally expensive and leads to faster learning [21].

The model (classifier) is then trained on training data drawn from raw dataset; raw dataset in current study refers to the dataset before its class imbalance was removed. Three classifiers used in current study are Decision Tree, Random Forests, LightGBM. The performance of the trained model is then measured on test dataset by measuring precision and F-score values. After calculating results on raw dataset, models are trained on randomly oversampled data. Random datapoints of minority class are replicated with replacement until their count matches the count of majority class datapoints which in current study is 284,315. Hence, total count of datapoints became 568,630 after oversampling. Training and test data are then drawn from this dataset and model is trained and tested on it. After random oversampling, experiments are repeated on dataset obtained after random under sampling. The total count of datapoints became 984 after random under sampling. Experiments were run on dataset obtained after SMOTE over-sampling. The precision and F-score values obtained for each model and data sampling technique are shown in Section 4.

Table 1. General form of Confusion Matrix

	Predicted: Yes	**Predicted: No**
Act-l: Yes	True positive (**TP**)	False negative (**FN**)
Act-l: No	False positive (**FP**)	True negative (**TN**)

4. Results and Discussion

Results obtained on decision tree classifier using three data sampling techniques are discussed in Section 4.1, results for Random Forests classifier are discussed in Section 4.2 followed by results for LightGBM in Section 4.3.

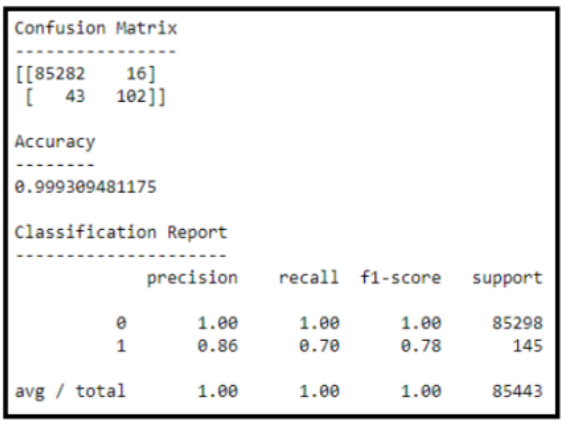
4.1. Decision Tree

Figure 1 shows the confusion matrix, test accuracy percentage, and classification report on test data obtained from raw dataset.

The confusion matrix values in Figure 1 can be understood by referring to Table 1. First row in the confusion matrix refers to count of non-fraudulent class and second row refers to count of fraudulent class, and all counts are with respect to test data.

The Figure 1 shows that out of total count of 85,298 records of nonfraudulent class in test dataset, 85,282 records were classified correctly and only 16 records were misclassified.

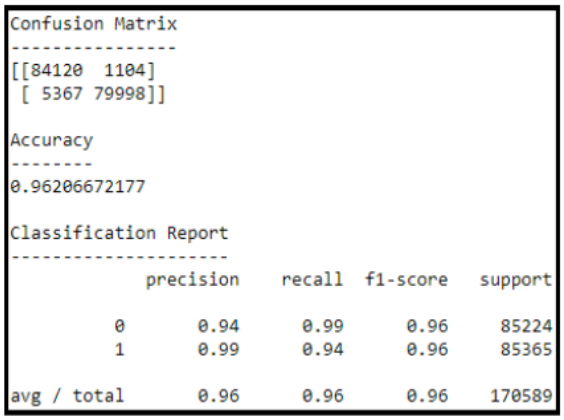
 | DOI: 10.20533/jitst.2046.3723.2021.0091



```
Confusion Matrix
----------------
[[85282    16]
 [   43   102]]

Accuracy
--------
0.999309481175

Classification Report
---------------------
             precision    recall  f1-score   support

          0       1.00      1.00      1.00     85298
          1       0.86      0.70      0.78       145

avg / total       1.00      1.00      1.00     85443
```

Figure 1. Decision Tree results on test data obtained from raw dataset

In classification report, first row depicts values for non-fraudulent class (class 0) and second row represents values for fraudulent class (class 1). Figure 2 rep-resents results for SMOTE oversampling. Out of all data sampling techniques, SMOTE oversampling achieved highest F-score for the fraudulent class. On comparing confusion matrix of SMOTE and the raw test data, it can be seen that false negative values are less in SMOTE and it achieves better performance.

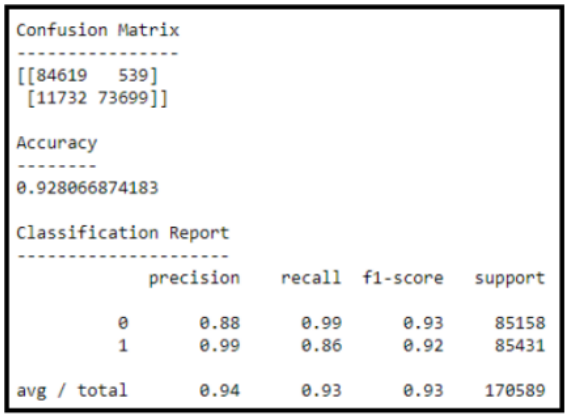


```
Confusion Matrix
----------------
[[84120  1104]
 [ 5367 79998]]

Accuracy
--------
0.96206672177

Classification Report
---------------------
             precision    recall  f1-score   support

          0       0.94      0.99      0.96     85224
          1       0.99      0.94      0.96     85365

avg / total       0.96      0.96      0.96    170589
```

Figure 2. Decision Tree results on test data obtained from SMOTE oversampled dataset

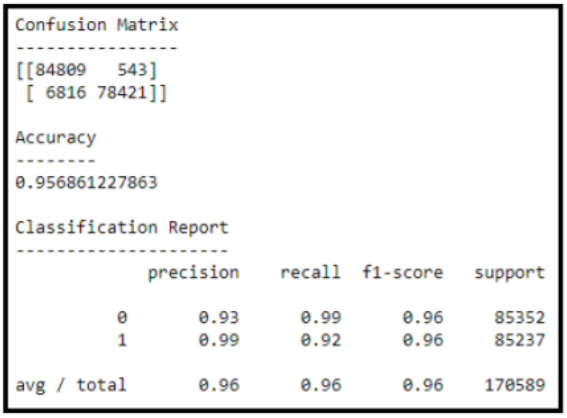
4.2. Random Forests

The Figure 3 shows confusion matrix, test accuracy percentage, and classification report on test data obtained from the raw dataset, when classified using Random Forests. The test datapoints in Random Forests for raw and sampled data is almost same as Decision Tree. Figure 4 represents SMOTE oversampling, when classified using Random Forests.



```
Confusion Matrix
----------------
[[84619   539]
 [11732 73699]]

Accuracy
--------
0.928066874183

Classification Report
---------------------
             precision    recall  f1-score   support

          0       0.88      0.99      0.93     85158
          1       0.99      0.86      0.92     85431

avg / total       0.94      0.93      0.93    170589
```

Figure 3. Random Forests results on test data obtained from raw dataset

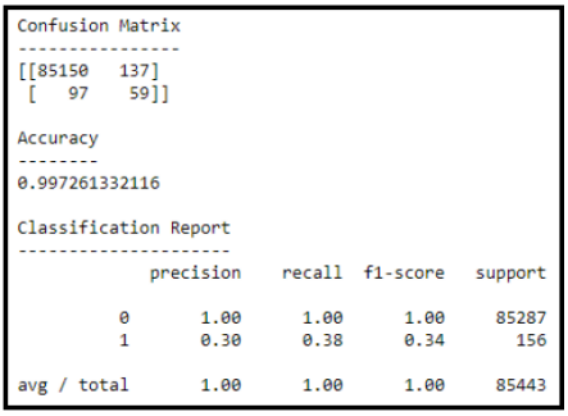


```
Confusion Matrix
----------------
[[84809   543]
 [ 6816 78421]]

Accuracy
--------
0.956861227863

Classification Report
---------------------
             precision    recall  f1-score   support

          0       0.93      0.99      0.96     85352
          1       0.99      0.92      0.96     85237

avg / total       0.96      0.96      0.96    170589
```

Figure 4. Random Forests results on test data obtained from SMOTE oversampled dataset

4.3. LightGBM

Figure 5 shows confusion matrix, test accuracy percentage, and classification report on test data obtained from raw dataset. The count of test datapoints in Random Forests for raw and sampled data is almost same as Decision Tree.

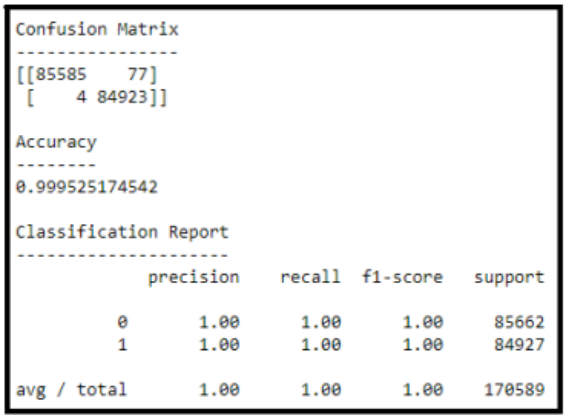


```
Confusion Matrix
----------------
[[85150   137]
 [   97    59]]

Accuracy
--------
0.997261332116

Classification Report
---------------------
             precision    recall  f1-score   support

          0       1.00      1.00      1.00     85287
          1       0.30      0.38      0.34       156

avg / total       1.00      1.00      1.00     85443
```

Figure 5. LightGBM results on test data obtained from raw dataset 10

 | DOI: 10.20533/jitst.2046.3723.2021.0091

Figures 6 represent results for SMOTE oversampling using LightGBM. It was noted that there were only four false positives in SMOTE oversampling.



```
Confusion Matrix
----------------
[[85585    77]
 [    4 84923]]

Accuracy
--------
0.999525174542

Classification Report
---------------------
             precision    recall  f1-score   support

          0       1.00      1.00      1.00     85662
          1       1.00      1.00      1.00     84927

avg / total       1.00      1.00      1.00    170589
```

Figure 6. LightGBM results on test data obtained from SMOTE oversampled dataset

5. Conclusion

The digital mode of payment is becoming more popular than ever before, and the risk of debit or credit card information being accessed by unintended people has increased. Hence, there is a need to have a subtle system which detects such anomalous activities as soon as they happen so that financial loss is reduced to a minimum. Such systems need to evolve over time so that intrusion techniques of intruders are nullified, and sensitive data of banks and customers is protected. This study is a contribution to improve the performance of existing systems. Experimental results showed that SMOTE oversampling achieved best precision and F-score values on three classifiers and out of three classifiers, LightGBM outperformed Decision Tree and Random Forests.

6. References

[1] Stephanie Chevalier. Global retail E-commerce sales 2014-2024. Jul 7, 2021. https://www.statista.com/statistics/379046/worldwide-retail-E-commerce-sales/. (Access Da -te: 29 November 2021).

[2] Financial Services Digital Fraud Attempts in Canada Rise 218%. TransUnion. https://www.transunion.ca/blog /fraudtrends-Q2-2021. (Access Date: 29 November 2021).

[3] Jain, Y., Namrata Tiwari, S. and Jain, S. (2019). A comparative analysis of various credit card fraud detection techniques. International Journal of Recent Technology and Engineering, 7(5S2), 402-407.

[4] Chawla, N. V., Japkowicz, N. and Kotcz. A. Editorial: Special issue on learning from imbalanced data sets. SIGKDD Explor. Newsl., 6(1):16, June 2004.

[5] Chan, P., Fan, W., Prodromidis, A. and Stolfo, S. (1999). Distributed data mining in credit card fraud detection. IEEE Intelligent Systems and their Applications, 14(6):67–74, 1999.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer. W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357.

[7] Chawla, N. V., Lazarevic, A., Hall, L. O. and Bowyer. K. W. (2003). Smoteboost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery, pages 107–119. Springer.

[8] Batuwita, R. and Palade, V. (2010). Efficient resampling methods for training support vector machines with imbalanced datasets. Proceedings of the International Joint Conference on Neural Networks, pp. 1–8.

[9] Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2006). Handling imbalanced datasets: A review," GETS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006. 12.

[10] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J. and Napolitano, A. (2008). RUSBoost: Improving classification performance when training data is skewed. 19th International Conference on Pattern Recognition, , no. IEEE, pp. 1–4.

[11] Guo, X. and Zhou, G. (2008). On the Class Imbalance Problem. Natural Computation, 2008. ICNC'08. Fourth International Conference. vol. 4, pp. 192–201.

[12] Chawla, N. V., Bowyer, K., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res., vol. 16, pp. 321–357.

[13] Gong, R., Fonseca, E., Bogdanov, D., Slizovskaia, O., Gomez, E. and Xavier Serra, (2017). Acoustic Scene Classification By Fusing Lightgbm and Vgg-Net Multichannel Predictions. IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events, no. November.

[14] Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE Trans. Syst. Man. Cybern. vol. 21, no. 3, pp. 660–674, 1991.

[15] Gislason, P. O., Benediktsson, J. A. and Sveinsson, J.R. (2006). Random forests for land cover classification. Pattern Recognit. Lett., vol. 27, no. 4, pp. 294–300.

[16] Liu, C. (2017). Everything You Need to Know about Alipay and WeChat Pay. Medium. https://charliecliu.medium .com/ everything-youneed-to-know-about-alipay-and-wechat-pay-2e5e6686d6dc. (Access Date: 30 November 2021).

[17] Wang, D., Lin, J., Cui, P., Jia, Q., Wang, Z., Fang, Y. and Qi, Y. (2019, November). A semi-supervised graph attentive network for financial fraud detection. In 2019

 | DOI: 10.20533/jitst.2046.3723.2021.0091

IEEE International Conference on Data Mining (ICDM) (pp. 598-607). IEEE. DOI: 10.1109/ICDM.2019.00070.

[18] Roa, L., Correa-Bahnsen, A., Suarez, G., Cortés-Tejada, F., Luque, M. A., andand Bravo, C. (2021). Super-app behavioral patterns in credit risk models: Financial, statistical and Journal of Internet Technology.

 | DOI: 10.20533/jitst.2046.3723.2021.0091