cek paper

by Yoana Nurul Asri, S.Si., M.Pd

Submission date: 01-Sep-2024 06:57AM (UTC-0700)

Submission ID: 2437055644

File name: CONFERENCE-PUBLISH-Data_Balance_Optimization_of_Fraud_Classification_for_E-Commerce_Transaction_1_.docx (218.29K)

Word count: 2614

Character count: 14732

Data Balance Optimization of Fraud Classification for E-Commerce Transaction

Aida Fitriyani
Informatics
Universitas Bhayangkara Jakarta Raya
Jakarta, Indonesia 17121
aida.fitriyani@dsn.ubharajaya.ac.id

Wowon Priatna
Informatics
Universitas Bhayangkara Jakarta Raya
Jakarta, Indonesia 17121
wowon.priatna@dsn.ubharajaya.ac.id

Tyastuti Sri Lestari
Informatics
Universitas Bhayangkara Jakarta Raya
Jakarta, Indonesia 17121
tyas@dsn.ubharajaya.ac.id

Dwipa Handayani
Informatics Universitas
Bhayangkara Jakarta Raya
Jakarta, Indonesia 17121
dwipa.handayani@dsn.ubharajaya.ac.id

TB Ai Munandar
Informatics
Universitas Bhayangkara Jakarta Raya
Jakarta, Indonesia 17121
dr.tb@dsn.ubharajaya.ac.id

Amri
Digital Businesses
Atma Luhur Saints and Business Institute
Bangka Belitung, Indonesia 33172
amri@atmaluhur.ac.id

2022 Seventh International Conference on Informatics and Computing (ICIC) | 979-8-3503-4571-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIC56845.2022.10007028

***Abstract* — the purpose of this study is to solve the problem of unbalanced data for prediction and classification of fraudulent E-Commerce transactions. Data from Digital Commerce 360 in 2015 showed that fraud occurred as much as 35% of total e-commerce transactions. Quoted from Bisnis.com, based on the 2017 Fraud Management Insight report, this percentage of fraud can reduce consumer confidence. One method for predicting fraud is machine learning. Fraud data does not have a balance between data that is not fraudulent, causing the classification to be biased. So it is necessary to balance the data using the SMOTE algorithm. the results of the data balancing will be classified as fraudulent transactions using the Support vector machine, K- Nearst Neighbor, Naïve Bayes and C45 algorithms.**

Keywords—Oversampling, Fraud E-Commerce SMOTE, K-Nearst Neighbor, Support Vector Machine, Naïve Bayes, C45.

I. INTRODUCTION

Indonesia has always been an attractive market for the growth of the e-commerce and online shopping market due to its large youth population and strong economic growth. the number of online shoppers in Indonesia is growing every year along with the growing number of e-commerce. the FT Confidential Report reports that online shoppers in Indonesia grew by 11 million in 2017 bringing the total to 35 million online shoppers (compared to 2015's total of 24 million).

Nevertheless, data from Digital Commerce 360 in 2015 showed that fraud occurred as much as 35% of total e-commerce transactions. Quoted from Bisnis.com, based on the 2017 Fraud Management Insight report, this percentage of fraud can reduce consumer confidence. In addition, the results of a survey conducted by Kaspersky Lab involving 26 countries around the world, revealed that as many as 26% of online consumers in Indonesia are victims of financial fraud. In Indonesia, the Directorate of Cyber Crime under the Criminal Investigation Agency of the Indonesian National Police stated that consumer losses caused by e-commerce fraud reached 2.2 billion rupiah. the e-commerce business itself is closely related to fraudulent actions. Based on DBS Insight Asia, the image of online shopping that is thick with fraudulent practices is the main reason why consumers are reluctant to shop online in Indonesia. One of the methods for predicting machine learning fraud for fraudulent E-Commerce transactions [1] and credit card fraud [2]. One type of machine learning algorithm is a classification algorithm [3][4].

The weakness of the classification algorithm is that the accuracy results will be biased if the target data or class is not balanced [5] and the characteristics of the data [6]. to overcome the data balance with the data oversampling algorithm using the SMOTE algorithm [7]. Research [8][9] uses the SMOTE algorithm to reduce the impact of data balance on credit card fraud datasets, uses SMOTE to overcome data rewards using 5 datasets from various applications [10].

Several studies have overcome the balance of data using the SMOTE algorithm before the classification process including research [11] performing data rewards before classification using the Support Vector Machine (SVM), K-Nearst Neighbor (KNN) [12], C45 for diabetes classification [13], performing rewards data before Sentiment analysis using Support Vector Machine and Naïve Bayes [14].

From the problems and related research, the purpose of this research will be to do oversampling before the E-Commerce fraudulent transaction data is classified. Oversampling class using SMOTE and continued with data classification using SVM, KNN, Naïve Bayes and C45 algorithms.

II. RESEARCH WORK

Data mining is a science in the field of data science that uses machine learning algorithms, various sectors of researchers use data mining for research materials [14]. Research [15] succeeded in increasing the prediction of employee performance using data mining so that they can make decisions. the results of the classification can be used as an analysis to detect fraud in the financial sector in India [16] and the classification can detect financial fraud in investments in the Canadian transaction market [17]. the result of an increase in the balancing of class data affects the

performance generated by the classification algorithm [18]. the results of oversampling data produce an accuracy value of 14% reducing errors in data classification where the value obtained by C45 is better than Bayesian, neural network and decision tree. While research [19] the results of the reduction in data rewards yield an accuracy rate of 97% for the Decision tree classification, which is better than the accuracy produced by naïve bayes.

III. METHODOLOGY

A. *Data*

The data used in this study is the E-Commerce public fraud detection dataset from https://www.kaggle.com datasets/akashanrastogi7767/ecommerce-fraud-data. the data obtained for the number of attributes is 9 (Nine) classified into 2 Fraud classes (True and False) with a total data of 167 records.

B. *Pre-Prosesing Data*

In pre-processing the data will go through two stages of the process, the first is scaling using the min max scaler algorithm [20]. this stage aims to overcome the problem of dataset intervals that are quite far away. this long interval will make it difficult for the Naïve Bayes, KNN, C45 and Support Vector Machine algorithms to train.

C. *Feture Selection*

The next stage is feature selection which is the process selecting relevant features that affect the classification results both in terms of the effectiveness and efficiency of the algorithm's work.

D. *Determining Data Balance*

This stage is to determine the balance of the data class oversampling with the SMOTE method [21], where the data will be reproduced by creating a new sample that has the same characteristics as the existing sample with the aim of balancing the amount of data for each class label.

E. *Classification*

At this stage is the classification of fraud detection of E- Commerce transactions using classification algorithms such as SVM, Naïve Bayes, KNN and C45.

F. *Classification test*

The method used in this process is the confusion matrix which is used to test the performance of the classification [22].

IV. RESULT AND DISCUSSION

The dataset of fraudulent E-Commerce transactions used is 167 records. the dataset must be free from noise and valid before the classification process is carried out with several prepared scenarios. the dataset must be in accordance with the design and requirements of the Naïve Bayes, C45, KNN and SVM algorithms free from dataset problems such as data intervals [23].

A. *Support Vector Machine*

After the pre-processing of the data has been carried out, the first classification algorithm experiment is the Support Vector Machine (SVM) algorithm. the SVM classification experiment was carried out before and after the data was balanced by the SMOTE algorithm. the data before and after the SMOTE process can be seen in Figure 1. the classification generated by SVM can be seen in table I.

Fig. 1. Visualisasi Before and After Over Sampling Data

The results of the smote can be seen in Figure 1 that the class/target was initially unbalanced after the Smote process was carried out to become balanced. Where before oversampling the target data for true = 39 and False = 69, after oversampling using SMOTE true becomes 68 and False becomes 68.

TABLE I. RESULTS OF SVM CLASSIFICATION

Algorithm	Recall	Precision	Accurracy	F1 Score
SVM	0.08	1.00	0.68	0.15
SVM + SMOTE	0.25	0.27	0.50	0.46

B. *Naïve Bayes*

The results of the Naïve Bayes classification before and after balancing the data using SMOTE are shown in table II.

TABLE II. NAÏVE BAYES CLASSIFICATION RESULTS

Algorithm	Recall	Precision	Accuracy	F1 Score
Naïve Bayes	0.08	0.50	0.65	0.14
Naïve Bayes + SMOTE	0.42	0.71	0.74	0.53

C. *K-Nearst Neighbor*

The results of the Naïve Bayes classification before and after balancing the data using SMOTE are shown in table III.

TABLE III. KNN CLASSIFICATION RESULTS

Algorithm	Recall	Precision	Accuracy	F1 Score
KNN	0.33	0.44	0.44	0.38
KNN + SMOTE	0.50	0.33	0.59	0.46

D. *Decision tree C45*

The results of Decision tree C45 classification before and after data balancing using SMOTE are shown in table IV.

TABLE IV. C45 CLASSIFICATION RESULTS

Algorithm	Recall	Precision	Accuracy	F1 Score
C45	0.08	0.17	0.53	0.11
C45 + SMOTE	0.25	0.33	0.56	0.29

E. *Classification Evaluation Results*

Hasil the results of all classifications that have been carried out after oversampling the data using SMOTE are continued by testing the performance confusion matrix of the classification generated by SVM, Naïve Bayes, KNN and C45 based on the value of Recall, Precision, Accuracy, F1 Score shown in table V.

TABLE V. CLASSIFICATION RESULTS

	SVM	SVM + SMOTE	NB	NB+ SMOTE	KNN	KNN+ SMOTE	C45	C45+ SMOTE
Recall	80%	25%	80%	42%	33%	50%	80%	25%
Presisi	100%	27%	50%	71%	44%	33%	17%	33%
akurasi	68%	50%	65%	74%	44%	59%	53%	56%
F1 Score	15%	46%	14%	53%	38%	46%	11%	29%

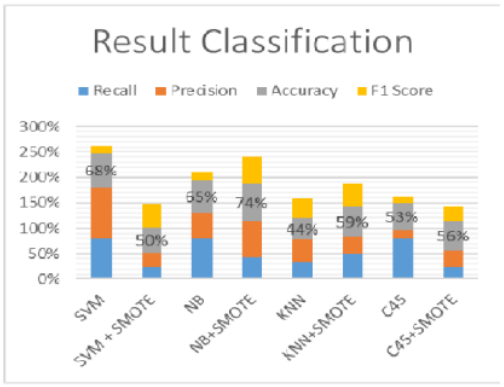

Fig. 2. Classification Visualization

From the results of fig 2 the results obtained before and after using SMOTE classification against the classification algorithm are:

- The classification results generated before using the SMOTE algorithm show that SVM has an accuracy of 68%, Naïve Bayes is 65%, KNN is 44% and C45 is 53%.
- The results of the classification after using the SMOTE algorithm are that SVM gets accuracy=50%, Naïve Bayes=74%, KNN=59%, C45=56%.
- The classification results generated before using the SMOTE algorithm showed that SVM got a recall of 80% Naïve Bayes by 80%, KNN by 33% and C45 by 80%.
- Classification results after using the SMOTE algorithm is that SVM gets Recall=25%, Naïve Bayes=42%, KNN=50%, C45=25%.
- The classification results generated before using the SMOTE algorithm show that SVM has a precision of 100% Naïve Bayes by 50%, KNN by 44% and C45 by 17%.
- Classification results after using the SMOTE algorithm are that SVM gets precision=27%, Naïve Bayes=71%, KNN=33%, C45=33%.
- The classification results generated before using the SMOTE algorithm showed that SVM got an F1 Score of 15% Naïve Bayes by 14%, KNN by 38% and C45 by 11%.
- The results of the classification after using the SMOTE algorithm were that SVM got F1 Score=46%, Naïve Bayes=53%, KNN=46%, C45=29%.
- Recommendations for the classification of E-Commerce fraud after oversampling using SMOTE is the Naïve Bayes algorithm with the highest accuracy value compared to the SVM, KNN and C45 algorithms.

V. CONCLUSION

The dataset used is e-commerce fraud which has an unbalanced class/target between the fraud class and the non- fraud class. to complete the data balance, the Synthetic Majority Over Sampling technique SMOTE algorithm is used. For the classification of the E-Commerce fraud dataset, 4 classification algorithms are used, including SVM, Naïve Bayes, KNN and C45. So the conclusions obtained in this research are:

- Before using SMOTE class Fraud=39 and non-fraud=69, after using the SMOTE class algorithm it becomes balanced with fraud being 68 and non-fraud being 68.
- The classification results generated before using the SMOTE algorithm show that SVM has an accuracy of 68%, Naïve Bayes is 65%, KNN is 44% and C45 is 53%.
- The results of the classification after using the SMOTE algorithm are that SVM gets accuracy=50%, Naïve Bayes=74%, KNN=59%, C45=56%.
- Recommendations for the classification of E-Commerce fraud after oversampling using SMOTE is the Naïve Bayes algorithm with the highest accuracy value compared to the SVM, KNN and C45 algorithms.

REFERENCES

[1] S. Carta, G. Fenu, D. Reforgiato Recupero, and R. Saia, "Fraud detection for E-commerce transactions by employing a prudential Multiple Consensus model," *J. Inf. Secur. Appl.*, vol. 46, pp. 13–22, 2019, doi: 10.1016/j.jisa.2019.02.007.

[2] V. N. Dornadula and S. Geetha, "Credit Card Fraud Detection using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 631–641, 2019, doi: 10.1016/j.procs.2020.01.057.

[3] M. Ito, K. Hoshino, R. takashima, M. Suzuki, M. Hashimoto, and H. Fujii, "Jou," *Healthc. Anal.*, p. 100119, 2022, doi: 10.1016/j.health.2022.100119.

[4] M. K. Severino and Y. Peng, "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real- world microdata," *Mach. Learn. with Appl.*, vol. 5, no. July 2020, p. 100074, 2021, doi: 10.1016/j.mlwa.2021.100074.

[5] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*,

vol. 23, no. 4, pp. 687–719, 2009, doi: 10.1142/S0218001409007326.
[6] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, "Review of Classification Methods on Unbalanced Data Sets," *IEEE Access*, vol. 9, pp. 64606–64628, 2021, doi: 10.1109/ACCESS.2021.3074243.
[7] A. Saputra and Suharjito, "Fraud detection using machine learning in e-commerce," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 9, pp. 332–339, 2019, doi: 10.14569/ijacsa.2019.0100943.
[8] Ri. Siringoringo, "Klasifikasi Data tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor," *J. ISD*, vol. 3, no. 1, pp. 44–49, 2018, [Online]. Available: https://ejournal-medan.uph.edu/index.php/isd/article/view/177/63.
[9] W. Nugraha, M. S. Maulana, and A. Sasongko, "Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012014.
[10] F. thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, 2020, doi: 10.1016/j.ins.2019.11.004.
[11] H. Ibrahim, S. A. Anwar, and M. I. Ahmad, "Classification of imbalanced data using support vector machine and rough set theory: A review," *J. Phys. Conf. Ser.*, vol. 1878, no. 1, 2021, doi: 10.1088/1742-6596/1878/1/012054.
[12] Z. Shi, "Improving k-Nearest Neighbors Algorithm for Imbalanced Data Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 719, no. 1, 2020, doi: 10.1088/1757-899X/719/1/012072.
[13] W. Nugraha and R. Sabaruddin, "Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4.5, Random Forest, dan SVM," *techno.Com*, vol. 20, no. 3, pp. 352–361, 2021, doi: 10.33633/tc.v20i3.4762.
[14] A. C. Flores, R. I. Icoy, C. F. Pena, and K. D. Gorro, "An evaluation of SVM and naive bayes with SMOTE on sentiment analysis data set," *ICEAST 2018 - 4th Int. Conf. Eng. Appl. Sci. technol. Explor. Innov. Solut. Smart Soc.*, pp. 1–4, 2018, doi: 10.1109/ICEAST.2018.8434401.
[15] M. Sharma and A. Goyal, "An application of data mining to improve personnel performance evaluation in higher education sector in India," *Conf. Proceeding - 2015 Int. Conf. Adv. Comput. Eng. Appl. ICACEA 2015*, pp. 559–564, 2015, doi: 10.1109/ICACEA.2015.7164755.
[16] S. Makki, "An Efficient Classification Model for Analyzing Skewed Data to Detect Frauds in the Financial Sector," no. 2019LYSE1339, 2019, [Online]. Available: https://tel.archives-ouvertes.fr/tel- 02457134.
[17] M. E. Lokanan and K. Sharma, "Fraud prediction using machine learning: the case of investment advisors in Canada," *Mach. Learn. with Appl.*, vol. 8, no. August 2021, p. 100269, 2022, doi: 10.1016/j.mlwa.2022.100269.
[18] A. Indrawati, H. Subagyo, A. Sihombing, W. Wagiyah, and S. Afandi, "Analyzing the Impact of Resampling Method for Imbalanced Data text in Indonesian Scientific Articles Categorization," *Baca J. Dokumentasi Dan Inf.*, vol. 41, no. 2, p. 133, 2020, doi: 10.14203/j.baca.v41i2.702.
[19] K. Kaur, "Credit Card Fraud Detection using Imbalance Resampling Method with Feature Selection," *Int. J. Adv. trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2061–2071, 2021, doi: 10.30534/ijatcse/2021/811032021.
[20] K. Bin Saboor, Q. Ul, A. Saboor, L. Han, and A. S. Zahid, "Predicting the Stock Market using Machine Learning: Long short-term Memory," *Electron. Res. J. Eng. Comput. Appl. Sci. www.erjsciences.info*, vol. 2, no. January 2021, p. 202, 2020, [Online]. Available: https://ssrn.com/abstract=3810128.
[21] D. Bajer, B. Zonc, M. Dudjak, and G. Martinovic, "Performance Analysis of SMOTE-based Oversampling techniques When Dealing with Data Imbalance," *Int. Conf. Syst. Signals, Image Process.*, vol. 2019-June, pp. 265–271, 2019, doi: 10.1109/IWSSIP.2019.8787306.
[22] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny).*, vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.
[23] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3446–3453, 2012, doi: 10.1016/j.eswa.2011.09.033.
[24] U. R. Gurning and Mustakim, "Penerapan Algoritma K-Means dan K-Medoid untuk Pengelompokkan Data Pasien Covid-19," *Build. Informatics, technol. Sci.*, vol. 3, no. 1, p. 48–55, 2021, doi: 10.47065/bits.v3i1.1003.

cek paper

ORIGINALITY REPORT

10%	8%	7%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repository.ubharajaya.ac.id Internet Source	3%
2	egrove.olemiss.edu Internet Source	1%
3	Christensen L.. "Hazardous and Industrial Waste Proceedings, 30th Mid-Atlantic Conference", CRC Press, 2019 Publication	1%
4	eprints.soas.ac.uk Internet Source	1%
5	www.bb-2.info Internet Source	1%
6	doi.org Internet Source	<1%
7	hdl.handle.net Internet Source	<1%
8	Morabito, Federico, Salvatore Nicosia, Andrew R. Teel, and Luca Zaccarian. "Measuring and Improving Performance in	<1%

	Anti-Windup Laws for Robot Manipulators", Springer Tracts in Advanced Robotics, 2004. Publication	
9	www.researchgate.net Internet Source	<1 %
10	Submitted to The WB National University of Juridical Sciences Student Paper	<1 %
11	www.wou.edu Internet Source	<1 %
12	research.tees.ac.uk Internet Source	<1 %
13	Aida Fitriyani, Wowon Priatna, Tyastuti Sri Lestari, Dwipa Handayani, TB Ai Munandar, Amri. "Data Balance Optimization of Fraud Classification for E-Commerce Transaction", 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022 Publication	<1 %
14	Riccardo Russo, Elaine Fox, Robert J. Bowles. "On the Status of Implicit Memory Bias in Anxiety", Cognition & Emotion, 7/1/1999 Publication	<1 %
15	de Solages, J. Baissus. "Greek synopsis of the gospels. (Engl. ed.)", Brill, 1959 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		