Jan 17th, 12:00 AM

Augmenting Data with Generative Adversarial Networks to Improve Machine Learning-Based Fraud Detection

Philipp Fukas
Osnabrück University, philipp.fukas@uni-osnabrueck.de

Lukas Menzel
Strategion GmbH, lumenzel@uni-osnabrueck.de

Oliver Thomas
German Research Center for Artificial Intelligence, oliver.thomas@dfki.de

Augmenting Data with Generative Adversarial Networks to Improve Machine Learning-Based Fraud Detection

Philipp Fukas[1 2 3], Lukas Menzel[1], and Oliver Thomas[1 2]

[1] Osnabrück University, Osnabrück, Germany
{philipp.fukas,lumenzel,oliver.thomas}@uni-osnabrueck.de
[2] German Research Center for Artificial Intelligence, Osnabrück, Germany
{philipp.fukas,oliver.thomas}@dfki.de
[3] Strategion GmbH, Osnabrück, Germany
philipp.fukas@strategion.de

Abstract. While current machine learning methods can detect financial fraud more effectively, they suffer from a common problem: dataset imbalance, i.e. there are substantially more non-fraud than fraud cases. In this paper, we propose the application of generative adversarial networks (GANs) to generate synthetic fraud cases on a dataset of public firms convicted by the United States Securities and Exchange Commission for accounting malpractice. This approach aims to increase the prediction accuracy of a downstream logit, support vector machine (SVM), and eXtreme Gradient Boosting (XGBoost) classifier by training on a more well-balanced dataset. While the results indicate that a state-of-the-art machine learning model like XGBoost can outperform previous fraud detection models on the same data, generating synthetic fraud cases before applying a machine learning model does not improve performance.

Keywords: Machine Learning, Fraud Detection, Data Augmentation, Generative Adversarial Networks, Financial Auditing

1 An Introduction to Machine Learning-Based Fraud Detection

To this date, financial fraud remains notoriously hard to detect and still requires a lot of manual forensic accounting work to be successfully uncovered. To reduce the amount of manual labor needed and to guide financial auditors, numerous artificial intelligence and machine learning (ML) methods such as decision forests or artificial neural networks have been applied to the fraud detection as well as the auditing domain [1–3]. But the difficulty of financial fraud detection still lies within the severe class imbalance that most real-world datasets are afflicted with because ML algorithms work best with large and equal observation amounts of each class to be predicted [4, 5]. To reduce dataset imbalance various methods such as synthetic minority oversampling technique (SMOTE), random undersampling (RUS), or random oversampling exist [4, 6]. Technically, GANs are not a direct method to handle imbalanced datasets, but their unique properties allow the generation of synthetic data for underrepresented classes [7]. Thus, GANs have recently shown promising results for reducing dataset imbalance

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

in credit card and telecom fraud detection [8, 9]. But so far, GANs have not been applied specifically to the domain of financial statement fraud and therefore, this paper provides the first empirical evidence of whether modern data augmentation methods like GANs can improve financial statement fraud detection performance.

2 Technical Setup of the Financial Fraud Detection Pipeline

In our fraud detection pipeline, two different discriminators operate (cf. Figure 1). The first is the discriminator within the GAN, which distinguishes whether the examples produced by the GAN's generator are real or synthetic fraud cases. The second is the classifier, whose task is to distinguish between fraudulent or non-fraudulent cases. The GAN is then trained to generate realistic fraud cases, which are merged with the training data. Our GAN implementation uses the open-source *CTGAN* package of Xu et al. [7].

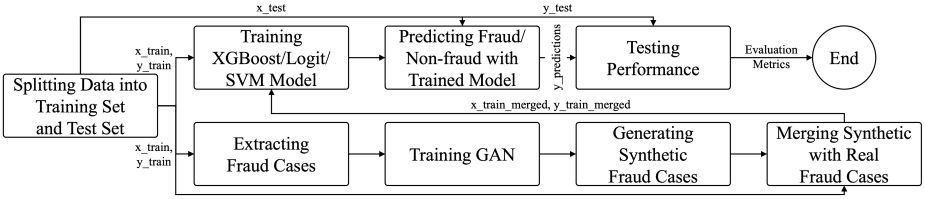

Figure 1. GAN and classifier training framework

To evaluate whether GANs can improve financial fraud detection performance, three logit models comparable to Dechow et al. [10], one SVM model corresponding to Cecchini et al. [11], and two state-of-the-art XGBoost models comparable to Bao et al. [12] were constructed. The in total six fraud detection models are trained on samples of publicly traded U.S. firms over the years 1991-2008. The raw data is based on publicly available financial data from financial statements featuring over 146.000 observations, but less than 1% of these are flagged as fraudulent firm-years [12]. Based on this raw data, three different sets of financial variables, that were previously identified by experts in fraud detection research, were created for the different models [10, 12]: (1) 28 raw data items used by one Logit, one SVM, and one XGBoost model, (2) a set of 14 ratio variables constructed from the raw data used by one logit model, and (3) the combined full set of 42 raw and ratio variables used by one logit and one XGBoost model. The classifiers are trained on the years 1991-2001 and tested on the years 2003-2008 to preserve the intertemporal nature of fraud detection. The GAN was also exclusively trained on all fraudulent firm-years in the period 1991-2001 to generate synthetic fraud cases. The model evaluation metrics area under the curve (AUC) [13] and normalized discounted cumulative gain (NDCG@k) [14] were also replicated from Bao et. al. [12]. For finding the optimal hyperparameters the classifiers are tuned using 3-fold cross-validation on a holdout validation set. The classifiers trained on the training set without adding synthetic fraud cases act as a baseline whereas the effect of the synthetic fraud cases is assessed by injecting fraud cases generated by the GAN trained on either 14, 28, or 42 variables into the real training data yielding in training datasets with a 1%, 2%, 5%, 10% and 20% fraud to non-fraud ratio.

3 Results of the Financial Fraud Detection Pipeline

Table 1. Out-of-sample performance metrics of the best performing models

Classifier	Training set fraud percentage	Performance Metrics on the test set 2003-2008			
		AUC	NDCG@k	Recall	Precision
XGB-28 (28 raw data items)	Baseline	0.761	0.050	5.36%	3.98%
	1%	0.604	0.014	1.53%	1.14%
	2%	0.621	0.025	2.68%	1.99%
	5%	0.660	0.021	2.30%	1.70%
	10%	0.580	0.021	2.30%	1.70%
	20%	0.589	0.004	0.38%	0.28%
XGB-42 (28 raw data items + 14 ratio variables)	Baseline	0.735	0.043	4.60%	3.41%
	1%	0.740	0.047	4.98%	3.98%
	2%	0.727	0.050	5.36%	3.98%
	5%	0.718	0.039	4.21%	3.12%
	10%	0.731	0.039	4.21%	3.12%
	20%	0.725	0.043	4.60%	3.41%

Our results show that all logit models outperform the SVM model. Most notably, our state-of-the-art XGBoost models outperform the RUSBoost model by Bao et al. [12] as well as all other logit and SVM classifiers. Therefore, only the results of the XGBoost models will be presented (cf. Table 1). The best performing XGBoost model trained on 28 raw financial variables has an AUC of 0.761 while the RUSBoost achieved an AUC of 0.725. NDCG@k is within a similar range at 0.050 (XGBoost) vs. 0.049 (RUSBoost). Likewise, the XGBoost algorithm trained on all 42 variables also outperforms previous attempts with an AUC of 0.735 vs. 0.696 and an NDCG@k of 0.043 vs. 0.035. Thus, employing the most recent ML models can indeed advance results for financial fraud prediction. Analogous to previous results, the addition of 14 financial ratios to the 28 raw data items does not induce a better classification performance [12]. Contrarily, there is no statistically significant evidence that augmenting the dataset with synthetic fraud cases generated by a GAN can improve the classification results. At any given synthetic fraud percentage, all classifiers either achieve similar or worse classification results when measured by AUC or NDCG@k. Notably, the classification performances of all six models are deteriorating with a rising percentage of generated fraud cases, suggesting a negative impact on the classifier.

4 Discussion and Conclusion

This paper provides the first empirical evidence of GANs applied specifically to the domain of financial statement fraud. While our XGBoost can outperform previous fraud detection models on the same data, the generation of synthetic fraud cases before applying a ML model conversely does not improve the performance. This might be because either the classifiers do not profit from the additional fraud cases or the GAN is not able to generate realistic fraud cases. Our further analyses with different similarity

metrics between the synthetic and the real data indicate, that the GAN is not fully able to learn the characteristics of a fraudulent firm-year and to generate realistic cases from it. But applying a different GAN implementation, which was able to generate more similar distributions, did not significantly improve our classification performance as well. Moreover, all previously mentioned results are limited to just one dataset of 42 variables. Training a GAN on a different dataset or different variable combinations might improve classification performance. This is especially important as the GAN used in this study is not perfectly capable of reproducing fraud cases with the same statistical properties. Further improving GANs for tabular data would be the most paramount future research objective, as GANs might not be fully capable of replicating the underlying statistical distribution yet. It is interesting, whether an improved generation of synthetic fraud cases will yield better classification results of a downstream classifier. Finally, a positive research outcome would improve the use of ML-based financial fraud detection pipelines significantly for the auditing practice.

References

1. Perols, J.: Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. Audit. A J. Pract. Theory. 30, 19–50 (2011).
2. Song, X.P., Hu, Z.H., Du, J.G., Sheng, Z.H.: Application of Machine Learning Methods to Risk Assessment of Financial Statement Fraud. J. Forecast. 33, 611–626 (2014).
3. Fukas, P., Rebstadt, J., Remark, F., Thomas, O.: Developing an Artificial Intelligence Maturity Model for Auditing. In: ECIS 2021 Research Papers (2021).
4. Kotsiantis, S., Koumanakos, E., Tzelepis, D., Tampakas, V.: Forecasting Fraudulent Financial Statements using Data Mining. Int. J. Comput. Intell. 3, 104–110 (2006).
5. Hasanin, T., Khoshgoftaar, T.M.: The Effects of Random Undersampling with Simulated Class Imbalance for Big Data. In: 19th International Conference on Information Reuse and Integration for Data Science. pp. 70–79. IEEE Press, Salt Lake City (2018).
6. Chawla, N. V.: Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O. and Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook. pp. 853–867. Springer, Boston (2005).
7. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular Data using Conditional GAN. In: Advances in Neural Information Processing Systems 32 (2019).
8. Fiore, U., De Santis, A., Perla, F., Zanetti, P., Palmieri, F.: Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. Inf. Sci. 479, 448–455 (2019).
9. Zheng, Y.J., Zhou, X.H., Sheng, W.G., Xue, Y., Chen, S.Y.: Generative adversarial network based telecom fraud detection at the receiving bank. Neural Networks. 102, 78–86 (2018).
10. Dechow, P.M., Ge, W., Larson, C.R., Sloan, R.G.: Predicting Material Accounting Misstatements. Contemp. Account. Res. 28, 17–82 (2011).
11. Cecchini, M., Aytug, H., Koehler, G.J., Pathak, P.: Detecting Management Fraud in Public Companies. Manage. Sci. 56, 1146–1160 (2010).
12. Bao, Y., Ke, B., Li, B., Yu, Y.J., Zhang, J.: Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. J. Account. Res. 58, 199–235 (2020).
13. Metz, C.E.: Basic Principles of ROC Analysis. Semin. Nucl. Med. 8, 283–298 (1978).
14. Järvelin, K., Kekäläinen, J.: Cumulated Gain-based Evaluation of IR Techniques. ACM Trans. Inf. Syst. 20, 422–446 (2002).