UNIVERSITI TEKNOLOGI MARA
Cawangan Kedah
Kampus Sungai Petani

2021 iCMS
INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e-PROCEEDINGS

of The 5th International Conference on Computing, Mathematics and Statistics (iCMS2021)

4-5 August 2021

Driving Research Towards Excellence

2021 iCMS

INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS

e-Proceedings of the 5th International Conference on Computing, Mathematics and Statistics (iCMS 2021)

Driving Research Towards Excellence

e-ISBN: 978-967-2948-12-4
DOI

Library of Congress Control Number:

The views and opinions and technical recommendations expressed by the contributors are entirely their own and do not necessarily reflect the views of the editors, the Faculty or the University.

Publication by
Department of Mathematical Sciences
Faculty of Computer & Mathematical Sciences
UiTM Kedah

TABLE OF CONTENT

PART 1: MATHEMATICS

PART 2: STATISTICS

PART 3: COMPUTER SCIENCE & INFORMATION TECHNOLOGY

PART 4: OTHERS

INVESTIGATING THE EFFECT OF DIFFERENT SAMPLING METHODS ON IMBALANCED DATASETS USING BANKRUPTCY PREDICTION MODEL

Amirah Hazwani Abdul Rahim[1] , Nurazlina Abdul Rashid[2], Abd-Razak Ahmad[3] and Norin Rahayu Shamsuddin[4]

[1,2,3,4]Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara (UiTM) Kedah Branch,
08400 Merbok, Kedah, Malaysia
([1]amirah017@uitm.edu.my,[2]azlina150@uitm.edu.my, [3]ara@uitm.edu.my, [4]norinrahayu@uitm.edu.my)

Abstract. Most classifiers of bankruptcy studies encounter less difficulty when dealing with a balanced non-bankrupt and bankrupt data set. The classifiers evaluate performance of the model through the accuracy rate. However, accuracy rate is not an appropriate measurement when dealing with imbalanced distribution of the data set. Sensitivity and precision were used instead to measure the performance of the classifier. This study employed three sampling strategies to deal with imbalanced datasets: oversampling, undersampling, and SMOTE (Synthetic Minority Oversampling Technique). The intent of this research is to examine how different sampling methods impact the performance of a bankruptcy prediction model utilising highly imbalanced real data. SMEs in the storage and transportation business were the subject of the research. The sample size is 9190 firms with 0.084% bankrupt firms and 99.16% non-bankrupt firms. As a classifier, Partial Least Square-Discriminant Analysis (PLS-DA) was selected. The findings suggest that employing Partial Least Square-Discriminant Analysis, SMOTE increases the classification probability for an imbalanced dataset. In the meantime, neither oversampling nor undersampling improved the results of the Partial Least Square-Discriminant Analysis.

Keywords- Partial Least Square-Discriminant Analysis, SMOTE, Oversampling, Undersampling, Imbalanced data

1. Introduction

The earliest study on financial distress prediction (FDP) begins with Beaver (1966) who used the univariate prediction model to show the significance of certain financial ratios in classifying bankrupt firms. Altman (1968) adopts multiple discriminant analysis (MDA) to build a multivariate model for FDP containing five financial measures as variables. The Z-score model is the product of the study. The Z-Score model, developed by Altman, is an analytical indicator that predict whether a firm will go insolvent in the following two years. Ohlson (1980) and Zmijewski (1984) utilised the Logit and Probit models, respectively, to get beyond the constraints of univariate investigation and MDA models. Kovacova and Kliestik (2017) found that a model based on logit functions slightly outperforms a probit model in terms of classification accuracy.

In the actual world of bankruptcy prediction, the proportion of bankrupt companies to non-bankrupt companies is not equal. It might be as minimal as 1 to 100 or as higher as 1 to 1000 (Veganzones and Séverin, 2018). The unsatisfactory efficiency of traditional classifiers, which algorithms are exclusively intended for balanced instances, has piqued curiosity in finding a solution to the issue of imbalanced datasets (Jia et al., 2014). The quantity of insolvent companies is much fewer than the total of companies that are not destitute, according to Zoriák et al. (2019). This aspect, however, is routinely overlooked in many articles, and balanced data is taken into account. Undoubtedly, bankruptcy prediction should consider this imbalance into account in hopes of avoiding type I and II faults, in which a company that is not bankrupt is rated as bankrupt, and vice versa.

Only some few research on the imbalance issue in bankruptcy prediction have been done eventhough the issues on imbalanced dataset has gained interest among researchers. Wilson and Sharda (1994), for contrast, compared neural networks with multivariate discriminant analysis using a resampling methodological framework, indicating that neural networks surpassed discriminant analysis in forecasting both non-bankrupt and bankrupt companies. In order to cope with the problem of an imbalance dataset, Japkowicz (2000) employed two resampling techniques: oversampling the minority class and downsizing the majority class. Since it only utilises a fraction of the majority class and is thus highly efficient, random under-sampling is a common approach for coping with class-imbalance concerns. Both oversampling the minority and downsizing and the majority were successful.

In another study, Zhou et al. (2012) examined the performance of more than 20 models for bankruptcy prediction using paired samples. In a prior work, Garcia et al. (2012) examined at the impact of the imbalance proportion and the classifier on the efficiency of four resampling techniques. When data sets are highly imbalanced, over-sampling of the minority class significantly outperforms over-sampling of the majority class, according to this study. On an actual heavily imbalanced dataset, Zhou (2013) investigated the productivity of an insolvency forecasting model utilising seven sampling approaches and five quantitative models. Model performance varies depending on the sampling methods used, however SVM can perform well in most situations.

Lin et al. (2017) apply two undersampling techniques to address the imbalance data sets as well as determine that undersampling the cluster centres' nearest neighbours is the best choice. Veganzones and Séverin (2018) explored different degree of imbalanced distributions through to recover from the loss of performance random oversampling, random undersampling, the synthetic minority oversampling technique (SMOTE), and EasyEnsemble. The outcome shows of their study shows that all the sampling technique achieve similar results on the recovery of performance loss. Nurazlina et al. (2017) conducted a study on PLS and logistic models for bankruptcy prediction model and found that the accuracy rate is close for both models. As a result, we would want to expand on the previous research by examining the impact of multiple sampling techniques on the success of bankruptcy prediction models and comparing the results of many frequently employed bankruptcy prediction models in this investigation.

2. Methodologies

2.1 Bankruptcy Data

The dataset is collected from Suruhanjaya Community Malaysia's Small and Medium Enterprises (SMEs) (SSM) from year 1999 to 2012. The data consists of 9113 non-failed and 77 failed Malaysian SMEs in the transportation and storage industry. SMEs are made up of variety of industries. This study solely looked at SMEs in the transportation and storage industry. As independent factors, financial ratios are employed. In the literary works, a wide variety of financial ratios were used to predict bankruptcy. Table 1 illustrates the financial ratios that were employed in this investigation.

Table 1: Financial Ratios

Label	Financial Ratio	Details
F1	NI/TA	Net Income/Total Assets
F2	CA/CL	Current Assets/Current Liabilities
F3	TL/TA	Total Liabilities/Total Assets
F4	WC/TA	Working Capital/Total Assets
F5	TL/TE	Total Liabilities/Total Equity
F6	S/TA	Sales/Total Assets
F7	CA/S	Current Assets/Sales
F8	CA/TA	Current Assets/Total Assets
F9	NI/S	Net Income/Sales
F10	NI/TE	Net Income/Total Equity
F11	TE/TA	Total Equity/Total Assets
F12	WC/S	Working Capital/Sales
F13	S/FA	Sales/Fix Assets
F14	TE/TL	Total Equity/Total Liabilities
F15	FA/TA	Fix Assets/Total Assets
F16	FA/TE	Fix Assets/Total Equity
F17	LTL/TA	Long-Term Liabilities/Total Assets
F18	CL/TA	Current Liabilities/Total Assets
F19	CL/TE	Current Liabilities/Total Equity
F20	EBT/TA	Earnings Before Taxes/Total Assets
F21	LTL/TE	Long-Term Liabilities/Total Equity
F22	S/TE	Sales/Total Equity
F23	TE/LTL	Total Equity/Long-Term Liabilities

2.2 Sampling strategies

The ‘Imbalanced’ package in R programming is used to analyze the sampling strategy. This study choose the common sampling technique which are random undersampling, random oversampling and SMOTE for imbalanced data. To achieve the balance that yields the equivalent result, random undersampling was used to eliminate the majority class. Oversampling, on the contrary, duplicates the minority class to attain the similar balance. SMOTE over-samples the minority class by producing synthetic minority instances in the vicinity of observed instances. The objective is to interpolate between examples of the same class to form new minority examples.

2.3 Performance Measures

The accuracy rate (Acc), sensitivity (Sen), specificity (Spec), and precision rate (Pre) are the four performance measures of the model.

Table 2: Confusion Matrix

	Predicted	
Actual Class	Positive (Bankrupt)	Negative(Non-Bankrupt)
Positive (Bankrupt)	True Positive (TP)	False Negative (FN)
Negative(Non-Bankrupt)	False Positive (FP)	True Negative (TN)

The accompanying estimations are drawn on Table 2

1. Sensitivity = $\frac{TP}{TP+FN}$
2. Specificity = $\frac{TN}{TN+FP}$
3. Accuracy Rate = $\frac{TP+TN}{TP+TN+FP+FN}$
4. Precision Rate = $\frac{TP}{TP+FP}$

3. Results

Training and testing samples were split from the main dataset. Table 3 shows that the original dataset in training for non- bankrupt and bankrupt cases is 5137 and 53 respectively. It indicates that the data was severely imbalanced. Following sampling, the minority to majority case class distribution for undersampling is 53:53, for oversampling it is 5137:5137, and for SMOTE it is 3975: 2703.

Table 3: Sampling method for imbalanced data

Sampling Methods	Training		Testing	
	Non-bankrupt	Bankrupt	Non-bankrupt	Bankrupt
	0	1	0	1
Original	5137	53	3976	24
SMOTE	3975	2703	3976	24
Undersampling	53	53	3976	24
Oversampling	5137	5137	3976	24

Table 4: Cross Validation for sampling methods

Sampling Methods	Actual Class	Predicted (Training)		Predicted (Testing)	
		Positive (Bankrupt)	Negative (Non-Bankrupt)	Positive (Bankrupt)	Negative (Non-Bankrupt)
Original	Positive (Bankrupt)	0	53	0	24
	Negative (Non-Bankrupt)	0	5137	0	3976
SMOTE	Positive (Bankrupt)	751	1952	6	18
	Negative (Non-Bankrupt)	321	3654	278	3698
Undersampling	Positive (Bankrupt)	32	21	11	13
	Negative (Non-Bankrupt)	17	36	1369	2607
Oversampling	Positive (Bankrupt)	2104	3033	7	17
	Negative (Non-Bankrupt)	812	4325	574	3402

Table 4 shows a cross validation for sampling methods using Partial Least Square Discriminant Analysis (PLS-DA) as a classifier. The result shows that for original sampling method positive and false positive are zero.

Table 5: Performance Measures

Sampling Methods	Training				Testing			
	Sensitivity	Specificity	Accuracy Rate	Precision Rate	Sensitivity	Specificity	Accuracy Rate	Precision Rate
Original	0.00	100.00	98.98	0.00	0.00	100.00	99.40	0.00
SMOTE	27.78	91.92	65.96	70.01	25.00	93.00	92.60	2.11
Undersampling	60.37	67.92	64.15	65.31	45.83	65.57	65.45	0.80
Oversampling	40.96	84.19	62.58	72.15	29.17	85.56	85.23	1.20

Table 5 demonstrates how different sampling techniques impact the bankruptcy prediction model's performance. As expected, the classifier is biased, with a high specificity (100%) and a sensitivity of 0%. As per the findings for SMOTE, undersampling, and oversampling, the trial set's sensitivity has risen to 25.0%, 29.17%, and 45.83%, respectively. Sensitivity, precision rate and accuracy rate of SMOTE sampling are better than oversampling and undersampling.

4. Conclusion and Discussion

In this experiment, we look at the impact of three different sampling methods on the categorization of an imbalanced dataset: undersampling, oversampling, and SMOTE sampling. SMOTE improves classification for imbalanced datasets by using Partial Least Square-Discriminant Analysis as a classifier, according to the findings of this study. Meanwhile, Oversampling and Undersampling did not improve the Partial Least Square-Discriminant Analysis performance.

References

Altman E (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance 23(4):589–609.

Beaver WH (1966). Financial ratios as predictors of failure. Journal of Accounting Research 4 (Supplement): 71–111.

Garcia V.,Sanchez J.S., Mollineda R.A. (2012), On the effectiveness of preprocessing methods when dealing with different levels of class imbalanced classification, Knowledge-Based System 25 : 3-12.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000):Special Track on Inductive Learning* Las Vegas, Nevada

Jia, Pengfei, Chunkai Zhang, and Zhenyu He. (2014). A New Sampling Approach for Classification of Imbalanced Data Sets with High Density. In 2014 International Conference on Big Data and Smart Computing (BIGCOMP), 217–222.

Kovacova, M. & Kliestik, T.(2017).Logit and Probit application for the prediction of bankruptcy in Slovak companies, Equilibrium. Quarterly Journal of Economics and Economic Policy, Institute of Economic Research, vol. 12(4), pages 775-791.

Lin, W.C.; Tsai, C.F.; Hu, Y.H.; Jhang, J.S. (2017) Clustering-based undersampling in class-imbalanced data. Inf. Sci. 409, 17–26

Ohlson JA (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18(1): 109–131.

Rashid N.A., Rahim A.H.A., Nasir IN.M., Hussin S., Ahmad AR. (2017) Classifying Bankruptcy of Small and Medium Sized Enterprises with Partial Least Square Discriminant Analysis. In: Ahmad AR., Kor L., Ahmad I., Idrus Z. (eds) Proceedings of the International Conference on Computing, Mathematics and Statistics (iCMS 2015). Springer, Singapore.

Veganzones, D., & Séverin, E. (2018). An investigation of bankruptcy prediction in imbalanced datasets. Decision Support Systems, 112, 111-124. https://doi.org/10.1016/j.dss.2018.06.011.

Wilson R., Sharda R. (1994), Bankruptcy prediction using neural network, Decision Support System 11: 545-557.

Zhou L., Lai K.K, Yen J. (2012), Empirical models based on features ranking techniques for corporate financial distress prediction, Computers & Mathematics with Application : 2484-2496.

Zhou, L. (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. Knowledge-Based Systems, 41, 16–25.

Zmijewski ME (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research 22: 59–82.

Zoričák, M., Gnip, P., Drotár, P., & Gazda, V. (2019). Bankruptcy prediction for small- and medium-sized companies using severely imbalanced datasets. Economic Modelling.

.

2021 ICMS

INTERNATIONAL CONFERENCE ON COMPUTING,
MATHEMATICS AND STATISTICS

e ISBN 978-967-2948-12-4

9 7 8 9 6 7 2 9 4 8 1 2 4