**Over or Under? Using sampling methods for Bank Failure Early Warning Systems**

Luke Artola

Columbia University

Master’s Thesis: Quantitative Methods in the Social Sciences

December 30th 2021

Supervisor: Benjamin Goodrich

**Acknowledgments**

First, I would like to thank the entire Quantitative Methods in the Social Sciences program for allowing me this wonderful opportunity to be a part of such a wonderful educational experience. The professors in the program helped to lay the educational foundation for this thesis and would have not been possible without their tutelage.

I would also like to thank my thesis advisor Prof. Benjamin Goodrich who has helped to guide me throughout this process for which I am eternally gratefully.

A special thank you has to be given to Prof. Mark Weinstock who is the reason I am in the position I am today. With his help I have been able to achieve more than I would have ever believed.

Lastly, I would like to thank my family for supporting me through my educational journey through the years.

**Abstract**

This paper looks to expand on the current literature on bank failure early warning systems regarding the effect of sample selection methods have on predictive ability. There has been very little work regarding optimal sample selection methods regarding bank failures. This is imperative as bankruptcy data is highly imbalanced causing regular statistical and machine learning methods to potentially have difficulty identifying failing banks with all the noise. Four sampling methods with a regularized logistic regression is applied to FDIC bank data to find the optimal sampling method to create a two year early warning system. **Continue with Results of Paper**

**Introduction**

Since the global financial Crisis (GFC) in 2008, there was 511 bank failures in the United States with the majority resulting from the crisis (FDIC). The GFC prompted governments across the world to enact new legislation and regulations aimed at promoting better stability for the financial system. As a result, Basel III was developed which are international supervisory guidelines meant to mitigate the risks financial institutions can pose due to lack of proper safe guards. In an effort to improve supervision there has been a resurgence in research regarding early warning systems (EWS) following the work of Beaver (1966) who originally used financial ratios to predict bank failures. These are statistical & machine learning models used by regulatory agencies to identify institutions that are potentially at risk of default. EWS allow regulators to proactively take action for at risk institutions minimizing the impact on consumers and the wider economy. These actions range from regulatory restrictions to assuming conservatorship of an institution.

Early warning systems have been created for a wide range of domains in an attempt to be better prepared for difficult situations. Researchers in China have created an air quality warning system for cities using six types of air pollution along with a support vector machine to help predict when pollution levels might become dangerously high (Xu, Yany, Wang 2017). In education EWS has been proven to help to identify potentially at-risk students allowing institutions to intervene earlier allowing a greater probability of success for those students. This study using Learning Management Data found that they could predict with 81% accuracy students who would receive a failing grade using logistic regression (Macfadyen, Dawson 2010). Political scientists have also developed an EWS for forecasting potential political violence in countries (Hegre et.al 2019). Using an Artificial Neural Network researchers in Turkey were able to correctly predict a currency crisis may occur within a 12-month period (Sevim, Oztekin, Bali, Gumus, Guresen 2014).

Many statistical methods have been used over the years in an attempt to quantify the likelihood of an institution failing. One popular method has been a Multivariate Logistic Regression (Martin, 1977; Ohlson, 1980). Most research regarding bank failures have been in the statistical methods domain. Within the past 15 years researchers have been focusing on the application of machine learning algorithms such as support vector machines (Erdogan, 2013; Gogas et al.,2018), random forest (Vuono, Michael 2019), and Neural Networks (López-Iturriaga et al., 2010; Constantin et al. 2018). There is still a debate as to whether statistical methods or machine learning methods are the optimal solution for regulators with (Jing, Zhongbo, Fang. 2018; Beutel, List, von Schweinitz, 2019) both providing evidence that statistical methods may still be the method of choice.

Real world bank failure data is highly imbalanced with some cases having failures representing 10% of the sample with other samples having it represent less then 1% of the overall data. Imbalanced data is a common phenomenon in many fields such as fraud detection, medical detection, and spam detection. This can pose a major problem when trying to predict the outcome especially as in most cases the minority class are the class of interest. Unless accounted for, models tend to be biased towards the majority class reducing the predictive power of the model. A model could have a 97% accuracy on imbalanced data and still fail to correctly predict a single minority class. The two main ways practitioners account for imbalance class are using cost sensitive approaches or sampling. Cost sensitive approaches is the process of assigning costs or weights to the classes which will cause the model to increase the cost of misidentifying the minority class. Sampling approaches attempt to change the class balance of the data set allowing the algorithm to better identify the trends in the minority class thus having better predictive accuracy.

In this paper I have opted to use the sampling to mediate the effect of severe class imbalance in this data set. I will be using under sampling (US), oversampling (OS), Synthetic Minority Over Sampling Technique (SMOTE), and Adaptive Synthetic Sampling (ADASYN) to analyze the different effects of sampling methods on classification of bank failures to create a two year early warning system. Those sampling techniques will be applied to a regularized Logistic Regression to identify the optimal process.

The remainder of the paper is as follows: Section 2: literature review. Section 3: is a discussion of the data. Section 4: explains the methodology of the paper. Section 5: Examines the validation & results, Section 6: Conclusion & Discussion about the methodologies and results. Additionally, limitations and the potential for further research will be touched upon.

**2. Literature Review**

**Bank Failure Literature:**

Most research on predicting bank failures has been based on using financial ratios instead of nominal values as they do not capture the potential impact based on size constraints (Beaver 1966). The ratios used primarily focus on the international regulatory structure **CAMELS** which stands for ***C*** Capital adequacy ***A*** Asset quality ***M*** Management ***E*** Earnings ***L*** Liquidity ***S***Sensitivity capturing idiosyncratic risk. Some studies add macroeconomic indicators attempting to capture the potential systemic risk (Betz, Oprica, Peltonen, & Sarlin, 2014; Mayes & Stremmel). There is still a division as to the significance of macroeconomic variables with (Halling & Hayden, 2006; Vuono 2019) finding lack of significance in prediction ability. There are many differing views on the most critical variables with some claiming capitalization, profitability, and asset quality (Poghosyan & Čihák, 2009) while (Mayes & Stremmel 2012) found that the leverage ratio plays the most pivotal role. Kerstein, Kozberg (2013) found that all six CAMELS categories play a big role in prediction.

Researchers have applied a multitude of techniques ranging from logistic regressions to artificial neural networks. Early research mainly focused using logistic regression to estimate the probabilities of bank default (Martin, Pifer 1970). Martin (1977) compared discriminant models and logistic regression finding that they performed likewise if the main goal was classification prediction on US banks. Chiaramonte et al. (2016) predicted with 76% accuracy with a three-year forecast utilizing a Discrete time proportional hazards model centering on the z score. Erdogan, Birsen Eygi (2013) discovered that support vector machines were an optimal solution on Turkish institutions. Studying a group of banks in Europe over a five-year period Messai & Gallali (2015) found that using an artificial neural network that they could successfully predict bank failures with a two-year lag. Very recent studies though have discovered that random forest preforms very similarly to artificial neural networks using a two year early warning system. They suggest that random forest might be a potential less computational intense alternative (Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis, 2020; Rustam & Saragih, (2018). There is still quite a void in random forest classification research in this area.

**Data Sampling Literature:**

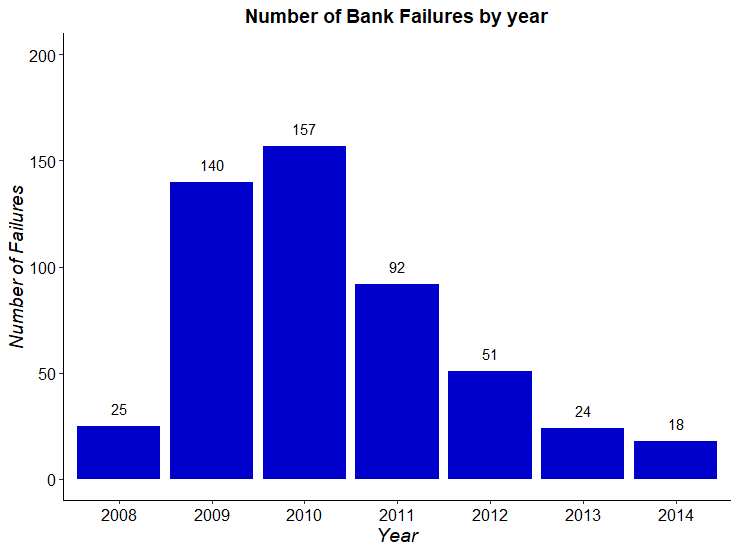
The major problem with predicting and forecasting bank failures is the large and sometimes quite severe class imbalances in the data. Most research in this domain has avoided directly dealing with the imbalance issue instead deciding to use paired samples based on size, regulator, or area of expertise. This sampling is done on the training and testing creating an artificial balance (Tam, Kar Yan, 1991; Ravisankar, Pediredla, and Vadlamani Ravi, 2010). This process helps to mitigate the effect of class bias which is introduced by the severe imbalance in the data. Along with that it reduces the computational costs associated with running the models. This leads to better classification rates but the problem with this is that a real world applied model will have to be able to identify failing banks where selecting paired samples is not a feasible solution. Bank failures tend to be a very small percentage of the total data set with some sets consistently being less then 1% of the data if not more.

There is very little literature regarding the class imbalances regarding bank failure prediction. Garcı ́a, Derrac, Triguero, Carmona, Herrera (2012) tested four resampling techniques with eight classification techniques on 17 different data sets. It was found that oversampling techniques provided the best classification accuracy on severely imbalanced data sets. Using a probit model Zmijewski (1984) found that a pairwise matching sample to predict financial distress caused a distortion of the probability of financial distress among observations. Neves & Vieria (2006) tested class distributions of 50:50, 36:64, 28:72 between solvent and insolvent French financial institutions. The more imbalanced samples caused a bias towards the healthier firms reducing the predictive ability of identify failing firms.

**3. Data**

The financial ratios which are to be used as predictor variables were collected from the FDIC’s Statistics of Depository Institutions database which contains all reported bank financial data received by their member institutions. The data used in this paper is quarterly based covering Q1 2006 - Q4 2012. The FDIC keeps a list of banks which have failed or have been in need of assistance. Due to the small sample size of failing banks, it had been decided to predict bank failures as well as banks in need of intervention and/or support. This collection of banks will all be referred to as insolvent throughout the rest of the paper. These banks have been encoded as a binary variable with 1 being an insolvent bank or bank requiring intervention and 0 being a solvent bank representing the dependent variable. The bank failures data was comprised of banks who were insolvent between the years of Q1 2008 – Q1 2014. The discrepancies between the dates of the financial data and insolvent banks were in an effort to create the two-year EWS so the financials are lagged two years. Given that the banks still existed in future quarterly statements due to the implementation of an EWS, I removed all financial information for insolvent banks past the two-year window to prevent potential noise. Following prior research, it was decided to drop the nominal values for average total assets, average earning assets, average equity, and average total loans as they would not be able to accurately take into account the relative size of the institution.

There were 9,307 N/A values present in the data which where all associated with the solvent class. Due to the large nature of the data set the author decided to remove the missing observations. Along with that there was a very large range of values for the variables in both directions. To mitigate the issue some but not all outliers were removed given the fact the minority class is very small so preservation of minority class observations was the primary objective. After the cleaning process there was 215,004 observations with 9,126 unique institutions. The minority class only represented 507 banks resulting in being 0.24% of the total sample. The largest amount of bank failures occurred in 2010 of 157 which constituted 1.89% of the banks in 2010.



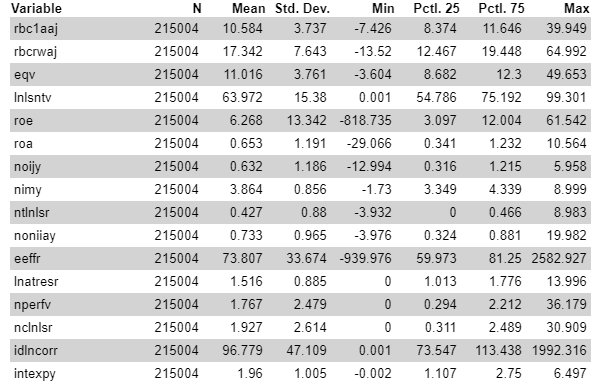
*Figure 1: Bank Failures by year. 2008 - 2014*

To obtain the training and testing sets, data from Q1 2008 - Q4 2012 were used while the observations from Q1 2013 - Q4 2014 were reserved for the out of time validation group. The train-test data had an 80/20 split stratified along the solvency indicator due to the highly imbalanced nature of the data. The testing data set had 126, 258 observations with 372 being identified as insolvent. This resulted in the insolvent class making up 0.29% of the observations. The testing set had 31, 657 observations with 93 sample being identified as insolvent making up 0.29%. A third data set containing the bank failures from Q1 2013 – Q4 2014 was created which represents an out of time sample. This data set will be used to gauge the model’s generalizability and will be the most closely looked at sample for how the model preforms as an early warning system. By incorporating the out-of-time sample it will be another robustness check for potential under or over fitting. There were 56,717 samples with 42 being insolvent resulting in being .07% of the observations. The final dataset resulted in having 17 variables including the solvency indicator. Due to the disagreement as to the superior set of variables to use, using Liu, Xian, Liu, Sathye (2021) literature review of bank failure prediction model papers I explored the most frequently used variables in papers to create a robust model.

**Variable Names & CAMELS Categorization**

|  |  |  |
| --- | --- | --- |
| Feature | Variable Name | CAMELS Category |
| Cost of Funding Earning Assets | intexpy | Earnings |
| Efficiency Ratio | eeffr | Management Capability |
| Equity to Asset Ratio | eqv | Capital Adequacy |
| Leverage Ratio | rbcrwaj | Capital Adequacy |
| Loan Lease Allowance to Loans | lnatresr | Asset Quality |
| Net Charge-offs to loans | ntlnlsr | Asset Quality |
| Net Loans and Leases to Total Assets | lnlsntv | Liquidity |
| Net Loans and Leases to Core Deposits | idlncorr | Liquidity |
| Net Interest Margin | nimy | Earnings |
| Net Operating Income to Assets | noijy | Management Capabilities |
| Non-Current Assets Plus Other Real Estate to Assets | nperfv | Asset Quality |
| Non-Current Loans to Loans | lnatresr | Asset Quality |
| Non-Interest Income to Average Assets | noniiay | Management Capabilities |
| Return on Assets | roa | Earnings |
| Return on Equity | roe | Earnings |
| Total Risk Based Capital | rbc1aaj | Liquidity |

**Summary Statistics**



**4. Methodology**

**Random Under Sampling:** The process in which the majority class is reduced at random tp achieve a desired ratio with the minority class. Usually resulting in a 1:1 ratio between the classes. Consider for every *ai* (*i = 1…n*) observation in the minority class randomly pick *bi* from the majority class until *ai = bi*. The training data had 372 observations from the solvent class with 372 observations from the insolvent class creating a 50:50 split.

**Over Sampling with Replacement:** This is the inverse of random under sampling but altered for feasibility. The minority class is sampled with replacement until the minority class achieves the desired balance usually 1:1. For every *ai* (*i = 1…n*) observation in the minority class pick at random *ai* (*i = 1…n*) and add it to the minority data set. Do this until the sample size is *ai* = *bi* with *b* representing the majority class.

**Random Over Sampling Examples (ROSE) :** ROSE increases the minority class by creating synthetic observations to achieve a desired ratio with the majority class. It fabricates synthetic samples from a kernel estimate of conditional density with a smoothing matrix.

The training set had 63,406 insolvent observations and 63,224 solvent observations.

**Synthetic Minority Oversampling Technique (SMOTE):** A observation is picked at random from the minority class. Using KNN the new synthetic observation is created in the feature space created K neighbors and place on a line segment connecting. K is usually defaulted to 5 and has been kept such as. SMOTE has two other tunning parameters that the author did attempt to tune to find the optimal balance within the data. Percent Over (OS) dictates how many synthetic samples should be created for the minority class. Percent Under (US) tunes how many extra cases are selected from the majority class in response to the creation of the synthetic observations. Tunning parameters of 100% and 200% percent over were tested. The arguments for percent under was 25%, 50%, 75%, 100%, 150%, and 200%.

**Logistic Regression:**

Binary Logistic regression is a generalized linear model that computes the log odds of the dependent variable. Due to the possibility of log odds not being constrained between 0-1 a logistic function is then applied which restricts values between 0-1 and creates a linear decision boundary unless additional extensions are added. Depending on where the probability cutoff is placed will determine the classification of the observation. If the probability cut off is 0.5 any value above 0.5 will be classified as a 1 while anything below would be a 0. Due to the imbalanced nature of the data, it is common for practitioners to lower the probability threshold to increase the rate of classification for the minority class.

Logistic Regression Equation:

logp1-p= 0+ 1x1+…+ Pxp

Elastic Net regularization will also be applied to the Logistic Regression. The purpose of elastic net regularization is to reduce the chance of the model overfitting the training data. Elastic Net uses L1 and L2 penalties to allow better generalizability of the model on unseen data.

**Elastic Net Equation:**

*L= ∑(ŷi-yi)2+ ∑2+  ∑||*

**5. Results & Validation**

**Sensitivity (Recall): TP/(TP/FN)**

Sensitivity measures how precisely the model is able to predict the positive class. This measure is as particular importance in this study due to the costs associated with potentially missing a failing bank.

**Specificity:**

Specificity calculates how accurately the true negatives are predicted correctly as compared those who the algorithm misclassified. This will help to inform as to how well the model predicts solvent banks

**Balanced Accuracy:**

Due to the imbalanced nature of this research classical accuracy would be very misleading metric. Classical accuracy measures the ability of a model to predict all observations in the sample disregarding potential class imbalance.

With an imbalanced data set this will cause accuracy to become artificially higher a much larger portion of the equation in true negatives & false positives. Balanced accuracy takes into account the class distribution of the samples by taking the average accuracy by class.

This measures how accurately the positive class (insolvent minority class) along with the negative class (solvent majority class) is classified.

**Area under the ROC Curve (AUC):**

The AUC is the area under the ROC Curve in which the curve represents the balance between Sensitivity and Specificity. It is impossible to have 100% Specificity or 100% sensitive as the further a metric is geared towards one it will have a more difficult time classifying the other. The AUC measures how well the classifier is at separating the two classes. The value ranges between 0 – 1 with 0.5 meaning that the classes are not separable and 1 meaning perfect separation between the classes. If the AUC is 0 then it is predicting the inverse values with a positive being considered a negative, likewise for the reverse. Values above 0.8 are considered to have good performance.

**Macro F-Meas (F1): (Only care about postives being picked correctly)**

F1 is the harmonic mean of the model’s precision and recall used to quantify the model’s ability to correct the positive class. F1 ranges between 0-1 with a perfect score being 1 meaning perfect precision and recall. Regular F1 uses the precision of both classes together calculating true positives over the sum of true and false positives. When dealing with as severe class imbalances this metric can cause a problem due to not taking the class distributions into account. Using F1 in this study would result in a very low score below 5 due to the fact there will always be a lot more false positives then true positives that will heavily skew the equation.

Macro F1 takes into account potential class imbalances by taking the recall and precision of each class, then averaging them.

**SMOTE 200 % Over Sample (Out of Time Sample)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Percent Under sample | Sensitivity | Specificity | Balanced Accuracy | AUC |
| 25 | 0.7381 | 0.9059 | 0.8220 | 0.9227 |
| 50 | 0.7619 | 0.9370 | 0.8494 | 0.9335 |
| 75 | 0.7381 | 0.9511 | 0.8446 | 0.9386 |
| 100 | 0.7143 | 0.9600 | 0.8372 | 0.9400 |
| 150 | 0.6667 | 0.9693 | 0.8200 | 0.9429 |
| 200 | 0.6429 | 0.9745 | 0.8087 | 0.9452 |

**SMOTE 100% Over Sample (Out of Time Sample)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Percent Under Sample | Sensitivity | Specificity | Balanced Accuracy | AUC |
| 25 | 0.8571 | 0.8964 | 0.8768 | 0.9377 |
| 50 | 0.8333 | 0.93 | 0.8812 | 0.9383 |
| 75 | 0.8095 | 0.9418 | 0.8757 | 0.9388 |
| 100 | 0.7619 | 0.9509 | 0.8564 | 0.9405 |
| 150 | 0.7619 | 0.9584 | 0.8601 | 0.9432 |
| 200 | 0.7619 | 0.9660 | 0.8640 | 0.9474 |

**Training Set**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sampling Method | Sensitivity | Specificity | Balanced Accuracy | AUC |
| Random Under sampling | 0.8495 | 0.8199 | 0.8347 | 0.9163 |
| Over Sampling with Replacement | 0.8608 | 0.8346 | 0.8477 | 0.9188 |
| ROSE | 0 |  |  | 0.88 |
| SMOTE 100 OS 25 US | 0.9906 | 0.5054 | 0.7480 | 0.95 |
| SMOTE 100 OS 50 US | 0.97 | 0.59 | 0.78 | 0.94 |
| SMOTE 100 OS 75 US | 0.95 | 0.74 | 0.84 | 0.95 |

**Test Set**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sampling Method | Sensitivity | Specificity | Balanced Accuracy | AUC |
| Random Under sampling | 0.7527 | 0.8401 | 0.7964 | 0.8829 |
| Over Sampling with Replacement | 0.7634 | 0.8341 | 0.7988 | 0.8843 |
| ROSE | 0.7849 | 0.8210 | 0.8030 | 0.88 |
| SMOTE 100 OS 25 US | 0.9677 | 0.4506 | 0.7092 | 0.88 |
| SMOTE 100 OS 50 US | 0.9462 | 0.5827 | 0.7645 | 0.88 |
| SMOTE 100 OS 75 US | 0.9032 | 0.6957 | 0.7995 | 0.88 |

**Out Of Time**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sampling Method | Sensitivity | Specificity | Balanced Accuracy | AUC |
| Random Under sampling | 0.7381 | 0.9684 | 0.8532 | 0.9421 |
| Over Sampling with Replacement | 0.7143 | 0.9669 | 0.8406 | 0.9423 |
| ROSE | 0.7143 | 0.9638 | 0.8391 | 0.94 |
| SMOTE 100 OS 25 US | 0.8571 | 0.8964 | 0.8768 | 0.9377 |
| SMOTE 100 OS 50 US | 0.8333 | 0.9290 | 0.8812 | 0.9383 |
| SMOTE 100 OS 75 US | 0.8095 | 0.9418 | 0.8757 | 0.9388 |

**6. Conclusion & Discussion**

For further research it would optimal to test the various sampling methods amongst different statistical and machine learning algorithms

**7. References**

Altini, (2015). Dealing with imbalanced data: undersampling, oversampling and proper cross-validation. https://www.marcoaltini.com/blog/dealing-with-imbalanced-data-undersampling-oversampling-and-proper-cross-validation

Blagus, Rok, and Lara Lusa.(2013). “Smote for High-Dimensional Class-Imbalanced Data.” *BMC Bioinformatics*, vol. 14, no. 1, https://doi.org/10.1186/1471-2105-14-106.

Beaver, William H. (1966). Financial ratios as predictors of failure. Journal of Accounting Research 1: 71–111.

Beutel, List, von Schweinitz, (2019). "Does machine learning help us predict banking crises?," Journal of Financial Stability, Elsevier, vol. 45(C).

Chiaramonte, Laura, Hong Liu, Federica Poli, and Mingming Zhou, (2016). How Accurately Can Z-score Predict Bank Failure? Financial Markets, Institutions & Instruments 25: 333–60.

Constantin, Andreea, Tuomas A. Peltonen, and Peter Sarlin. (2018). Network linkages to predict bank distress. Journal of Financial Stability 35: 226–41.

Cuneyt Sevim, Asil Oztekin, Ozkan Bali, Serkan Gumus, Erkam Guresen, (2014), Developing an early warning system to predict currency crises, European Journal of Operational Research, Volume 237, Issue 3, Pages 1095-1104, ISSN 0377-2217

Daniel Martin. (1977). Early warning of bank failure: A logit regression approach, Journal of Banking & Finance, Volume 1, Issue 3, , Pages 249-276.

Erdogan, Birsen Eygi. (2013). Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. Journal of Statistical Computation and Simulation 83: 1543–555.

Frank Betz, Silviu Oprică, Tuomas A. Peltonen, Peter Sarlin. (2014). Predicting distress in European banks, Journal of Banking & Finance, Volume 45

S. Garcı ́a, J. Derrac, I. Triguero, C.J. Carmona, F. Herrera, Evolutionary-basedselection of generalized instances for imbalanced classification, Knowledge-Based Systems 25 (2012) 3–12.

G. Karatas, O. Demir and O. K. Sahingoz, "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset," in IEEE Access, vol. 8, pp. 32150-32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

G.L. Kaminsky, S. Lizondo, C.M. Reinhart. (1998). The leading indicators of currency crises, IMF Staff Pap., 45, pp. 1-48.

Gogas, Periklis, Theophilos Papadimitriou, and Anna Agrapetidou. (2018). Forecasting bank failures and stress testing: A machine learning approach. International Journal of Forecasting 34: 440–55.

Halling, Michael and Hayden, Evelyn, Bank Failure Prediction.(May 2006).: A Two-Step Survival Time Approach

Håvard Hegre, Marie Allansson. “Views: A Political Violence Early-Warning System - Håvard Hegre, Marie Allansson, Matthias Basedau, Michael Colaresi, Mihai Croicu, Hanne Fjelde, Frederick Hoyles, Lisa Hultman, Stina Högbladh, Remco Jansen, Naima Mouhleb, Sayyed Auwn Muhammad, Desirée Nilsson, Håvard Mokleiv Nygård, Gudlaug Olafsdottir, Kristina Petrova, David Randahl, Espen Geelmuyden Rød, Gerald Schneider, Nina Von Uexkull, Jonas Vestby, (2019).” ViEWS: A political violence early-warning system”, SAGE Journals, , <https://journals.sagepub.com/doi/full/10.1177/0022343319823860>.

Jing, Zhongbo, and Yi Fang. 2018. Predicting US bank failures: A comparison of logit and data mining models. Journal of Forecasting 37:235–56.

L. Breiman. (2001). “Random forests,” in Machine Learning, vol. 45, pp. 5-32,.

Leah P. Macfadyen, Shane Dawson, (2010),Mining LMS data to develop an “early warning system” for educators: A proof of concept, Computers & Education, Volume 54, Issue 2, Pages 588-599, ISSN 0360-1315

Liu, Xian, Liu, Sathye (2021). Predicting Bank Failures: A Synthesis of Literature and Directions for Future Research. Journal of Risk and Financial Management 14: 474. https://doi.org/10.3390/jrfm14100474

López-Iturriaga, López-de-Foronda, & Pastor-Sanz. (2010). Predicting Bankruptcy Using Neural Networks in the Current Financial Crisis: A Study of US Commercial Banks.

Martin, Daniel. 1977. Early warning of bank failure: A logit regression approach. Journal of Banking & Finance 1: 249–76.

Mayes, David G. and Stremmel, Hanno. (2012). The Effectiveness of Capital Adequacy Measures in Predicting Bank Distress. Financial Markets & Corporate Governance Conference

Meyer, Pifer (1970). Prediction of bank failures, Journal of Finance, pp. 853-868.

Messai, A. S., & Gallali, M. I. (2015). Financial leading indicators of banking distress: A micro prudential approach-evidence from Europe. Asian Social Science, 11(21), 78.

J. Neves, A. Vieira, Improving bankruptcy prediction with Hidden LayerLearning Vector Quantization, European Accounting Review 15 (2006) 253–271

Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), 109–131.

Petropoulos, Siakoulis, Stavroulakis, & Vlachogiannakis, 2020. "Predicting bank insolvencies using machine learning techniques," International Journal of Forecasting, Elsevier, vol. 36(3), pages 1092-1113.

Poghosyan, T., & Čihák, M. (2009). Distress in European banks: An analysis based on a new dataset. IMF working papers. (pp. 1–37).

R. C. Bhagat and S. S. Patil, "Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest.(2015). IEEE International Advance Computing Conference (IACC), 2015, pp. 403-408, doi: 10.1109/IADCC.2015.7154739. <https://ieeexplore.ieee.org/abstract/document/7154739>

Tam, Kar Yan (1991). Neural network models and the prediction of bank bankruptcy. Omega 19: 429–45.

Ravisankar, Pediredla, and Vadlamani Ravi. 2010. Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP. Knowledge-Based Systems 23: 823–31.

Rustam, Z., & Saragih, G.S. (2018). Predicting Bank Financial Failures using Random Forest. 2018 International Workshop on Big Data and Information Security (IWBIS), 81-86.

Vuono, Michael.(2019). “Predicting Bank Insolvency with Random Forest Classification.” Predicting Bank Insolvency with Random Forest Classification, https://lup.lub.lu.se/student-papers/search/publication/8982037.

Zmijewski, (1984). “Methodological Issues Related to the Estimation of Financial Distress Prediction Models.” Journal of Accounting Research 22: 59–82. https://doi.org/10.2307/2490859.

K.S. Shin, K.J. Lee, H.J. Kim

Support vector machines approach to pattern detection in bankruptcy prediction and its contingency

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2004), pp. 1254-1259

References:

# Macro Economic Approach to bank Failure

https://pdfs.semanticscholar.org/28f0/39c903741b393c3631dc501871add6741aad.pdf?\_ga=2.247283463.267459489.1637302820-284921495.1637302820

# Smote approach to balancing data for RF on multiclass which they made into binary classes

R. C. Bhagat and S. S. Patil, "Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest," *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 403-408, doi: 10.1109/IADCC.2015.7154739.

<https://ieeexplore.ieee.org/abstract/document/7154739>

#  Our study clearly demonstrates the need to apply at least some sampling to big data with class imbalance and suggests the 50:50 class distribution does not produce the best Medicare fraud detection results.

R. Bauder and T. Khoshgoftaar, "Medicare Fraud Detection Using Random Forest with Class Imbalanced Big Data," *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 80-87, doi: 10.1109/IRI.2018.00019.

#Experimental results demonstrated that the proposed approach considerably increases the detection rate for rarely encountered intrusions.

G. Karatas, O. Demir and O. K. Sahingoz, "Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset," in *IEEE Access*, vol. 8, pp. 32150-32162, 2020, doi: 10.1109/ACCESS.2020.2973219.

# Logit model explained

https://reader.elsevier.com/reader/sd/pii/037842667790022X?token=4687AACA31E2440897EE0C442C5D965C0F0A51E46B1FA1D4F26CFEA71AAFD18B301B5658EF99740CFAB01FB3EABC45C9&originRegion=us-east-1&originCreation=20211217065001

Martin

# Smote for Indian Banks

Santosh Shrivastava, P Mary Jeyanthi & Sarbjit Singh | (2020) Failure prediction of Indian Banks using SMOTE, Lasso regression, bagging and boosting, Cogent Economics & Finance, 8:1, 1729569, DOI: 10.1080/23322039.2020.1729569

# Sampling techniques for non-financial corporate bankrupcites2009 and prior

https://reader.elsevier.com/reader/sd/pii/S095070511200353X?token=892DDB065C3C8073B23ECA5E46E200263844F99BAAB55BF036E5CD0437EE345804F51BA386D7FF901457FB898533A853&originRegion=us-east-1&originCreation=20211217072342

 M Gruszczyński · 2019

Sampling Bank failures in europe

file:///C:/Users/amaut/Downloads/ijfs-07-00028%20(1).pdf

Tam, Kar Yan. 1991. Neural network models and the prediction of bank bankruptcy. Omega 19: 429–45.

Ravisankar, Pediredla, and Vadlamani Ravi. 2010. Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP. Knowledge-Based Systems 23: 823–31.

(Tam, Kar Yan, 1991; Ravisankar, Pediredla, and Vadlamani Ravi, 2010)