**Machine Learning Applications for Predicting Bank Insolvencies**

Luke Artola

Columbia University

Masters Thesis: Quantitative Methods in the Social Sciences

Random Date

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.

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 regarding predicting bank failures using an early warning system. In comparison to other methodologies on this subject very little attention has been given until recently regarding the use of Random Forest. Along with that there is an even bigger gap in the current literature regarding the potential of using sampling methods in conjunction with early warning systems. In this paper I propose using sampling methods in combination with Random Forest to create a 2 year early warning system. **Continue with Results of Paper**

**Table of Contents**

**No table of contents entries found.**

**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 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. These actions range from regulatory restrictions to assuming conservatorship of an institution.

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). This method was found to not be reliable due to the linear assumptions made by logistic regression. It prevents the model from potentially capturing non-linear patterns. Due to this researcher’s began applying machine learning algorithms such as support vector machines (Erdogan, 2013; Gogas et al.,2018), Neural Networks (López-Iturriaga et al., 2010; . Constantin et al. 2018),

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

**2. Literature Review**

Predicting bank failures has been an extensive area of research for academics and regulatory bodies. Early research focused on inferential understanding of the causes which has evolved over the decades to be primarily concerned with the ability to construct accurate EWS. 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.

**3. Data**

The financial data 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 were two years prior to insolvency. Given that the banks still existed in future quarterly statements due to the lag, 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 BLANK as they would not be able to accurately take into account the relative size of the institution

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 group. The train-test data had an 80/20 split stratified along the solvency indicator due to the highly imbalanced nature of the data.

**4. Model & Model Development**

**5. Validation**

**6. Conclusion & Discussion**

**7. References**

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.

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

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,

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

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

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

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

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

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

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.

References:

# Original Random Forest Paper

# 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.