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

Advisor: Benjamin Goodrich

Acknowledgments

First, I would like to thank the Quantitative Methods in the Social Sciences program for the wonderful opportunity to be a part of such an incredible educational journey. The professors in the program helped lay the educational foundation for this thesis, as such the paper would have not been possible without their tutelage.

I would also like to thank my thesis advisor Prof. Benjamin Goodrich who has played a crucial role in guiding me throughout this process for which I am eternally grateful.

Additionally, 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 and encouragement, I have been able to achieve more than I would have ever believed.

Lastly, I would like to thank my family for all of their support throughout my educational journey through the years.

Abstract

This paper seeks to expand upon the current literature on bank failure early warning systems. Particularly regarding the effect sample selection methods have on predictive ability. There has been a lack of research regarding optimal sample selection methods when it comes to bank failures. This is imperative as bankruptcy data is highly imbalanced causing statistical and machine learning methods to have difficulty identifying failing banks due to class bias. Four sampling methods with a Regularized Logistic Regression were applied to FDIC bank call reports to find the optimal sampling method to create a two year early warning system. The data covered the periods from Q1 2008 – Q4 2014 with the periods Q1 2013- Q4 2014 providing a third Out of Time data set to test the practical application of the model. The research suggests that SMOTE is the optimal sampling method achieving a balanced accuracy of 87.5% on the Out of Time validation set.

1.Introduction

Since the Global Financial Crisis (GFC) in 2008, there have been 536 bank failures in the United States with the majority occurring in the following two years, 2009 with 140 failures and 2010 with 157 failures. The GFC prompted governments across the world to enact new legislation and regulations aimed at promoting stronger stability for the financial system. As a result, the international regulatory accord Basel III was developed which are supervisory guidelines meant to mitigate the risks financial institutions within the banking sector can pose due to lack of proper safeguards. 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. EWS's are statistical and machine learning models used by regulatory agencies to identify institutions that are potentially at risk of default. EWS allows 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 to encompass wide range of domains in an attempt to be better prepared for difficult situations ahead of time. For instance, in health sector 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 predict when pollution levels might become dangerously high (Xu, Yany, Wang 2017). This allows citizens to take preemptive action in purchasing or wearing masks that will filter the smog or pollution during bad air quality days. In education, EWS has been proven to help identify potentially at-risk students allowing institutions to intervene earlier allowing a greater probability of success for those students. For example, a 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). Likewise, 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 which has been used in several studies (Martin, 1977; Ohlson, 1980). While most research regarding bank failures has been in the statistical methods domain, within the past 15 years however, 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) providing evidence that statistical methods may still be the method of choice.

Real world bank failure data is highly imbalanced which means that for a two-class data set one class has the majority of observations. With some cases of bank failure data sets have failures representing 5% of the sample while other samples may have failures representing less than 1% of the overall data. Imbalanced data is a common phenomenon in many fields such as fraud detection, medical detection, and spam detection. 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. This can pose a major problem to researchers and practitioners when trying to predict the

outcome that generalizes multiple group trends, especially as in most cases the minority class are the class of interest. The two main ways practitioners account for imbalanced classes are using cost sensitive approaches or sampling. A cost-sensitive approach is the process of increasing costs or changing weights of the classes which will cause the model to increase the cost of misidentifying the minority class. Due to the higher cost associated with the minority class in the loss function the model will focus more closely on identifying those observations to minimize the error. On the other hand, sampling methods focus on changing the class balance of the sample towards an optimal distribution.

In this paper I have opted to use the sampling method only on the training data to mediate the effect of severe class imbalance in this data set enabling the model to be used for realistic forecasting. I will be using Random Under Sampling (US), Random Oversampling with Replacement (OS), Random Over Sampling Example (ROSE), and Synthetic Minority Over Sampling Technique (SMOTE) to analyze the different effects of sampling methods on classification of bank failures to create a two year early warning system on US banks from years 2008 to 2014. Those sampling techniques will be applied to a Penalized Logistic Regression with Elastic Net Parameterization to identify the optimal process.

The remainder of the paper is as follows will be section 2: reviewing the previous literature, section 3: discussing the data and cleaning, section 4: explaining the methodology of the paper, section 5: examining the validation and results, lastly section 6: discussing the results with a conclusion on the methodologies and outcome. 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 nominal values do not capture the potential impact based on size constraints (Beaver 1966). The ratios used in this paper 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). However, there is still a division in the scientific community 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, other deeming it to be asset quality (Poghosyan & Čihák, 2009) while (Mayes & Stremmel 2012) found that the leverage ratio plays the most pivotal role. Interestingly 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 on OLS Linear regression and Logistic regression to estimate the chance of default in addition to identifying the main drivers (Martin, Pifer 1970). For instance, 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 a three-year forecast utilizing a Discrete Time Proportional Hazards model centering on the z score. They were able to predict bank failures

76% of the time on US banks from 2004-2012. Erdogan, Birsen Eygi (2013) discovered that Support Vector Machines were a viable solution for Turkish institutions. Following up on their work Gogas et al. (2018) achieved a 98% accuracy with a Support Vector Machine. Studying a group of banks in Europe over a five-year period Messai & Gallali (2015) found that using an Artificial Neural Network they could successfully predict bank failures with a two-year lag 79.6% of the time. Non-performing loans were concluded as the best indicator of financial distress during expansionary time. Recent studies though have discovered that Random Forest performs very similarly to Artificial Neural Networks using a two year early warning system (Petropoulos, et al, 2020; Rustam & Saragih, (2018). They suggest that Random Forest might be a potentially less computational intense alternative. Petropoulos, et al (2020) performed a Random Under Sampling on their training set to achieve a 1:10 ratio of the minority class. This method resulted in Random Forest performing better on the testing set but slightly worse than the Neural Network on the Out of Time sample.

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 business. 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 sampling may reduce the computational costs associated with running the models depending on the sampling method of choice. While this leads to better classification

rates the critical problem with this is in a real world applied model it will have to identify failing banks in large unbalanced data so selecting paired samples is not a feasible solution for practical application. Bank failures tend to be a very small percentage of the total data set with some sets consistently being less than 1% of the data if not more.

Moreover, there is very little literature discussing the class imbalances that concern bank failure prediction. García, Derrac, Triguero, Carmona, Herrera (2012) tested four resampling techniques with eight classification techniques on 17 different types of data sets. It was found that oversampling techniques provided the best classification accuracy on severely imbalanced data sets. Karatas, Demir & Sahingoz (2020) used Synthetic Minority Over-sampling Technique (SMOTE) to counterbalance the imbalance of data from Intrusion Detection Systems which identifies attacks on computer networks. They found that sampling greatly improves the predictive accuracy of classification algorithms on highly imbalanced data. 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 insolvent and solvent French financial institutions. The more imbalanced samples caused a bias towards the healthier firms reducing the predictive ability of identifying failing firms. Zhou (2013) used six different sampling methods with five modeling techniques to predict corporate non-financial bankruptcies. He identified that undersampling techniques were more effective when there were hundreds of samples in the minority class. Over sampling in particular SMOTE was better at forecasting when there were less than 100 observations for the minority class. Shrivastava, Jeyanthi & Singh (2020) forecasted bank failures in India from 2000-2017 using bank specific, macroeconomic, and

market structure variables. They applied SMOTE to rebalance the data to then apply Lasso for feature selection. They were able to attain a Type II error rate of 64.34%.

When dealing with imbalanced data in other areas of research it is common practice to explore the potential application of sampling methods. Nevertheless, there appears to be a void in the research concerning this topic. The few papers that do use a sampling method in predicting bank failures tend to pick a singular sampling method to test. There is not one method that is the consistent throughout the different papers with researchers picking different over and under sampling methods. To my knowledge there is no research that examines how different sampling methods may compare in prediction accuracy. Sampling methods have the potential to significantly reduce the computal costs in applied forecasting as Artificial Neural Networks and Random Forest have emerged as popular areas of research. This may also lead to better overall accuracy of identification of insolvent banks as it would reduce or eliminate the class bias poised by highly imbalanced data sets.

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 in the US. The financial data used in this paper is quarterly based covering Q1 2006 - Q4 2012. The FDIC additionally keeps a list of banks which have failed. These banks have been encoded as a binary variable with 1 being an insolvent bank, 0 being a solvent bank representing the dependent variable. The bank failures data was composed 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 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.

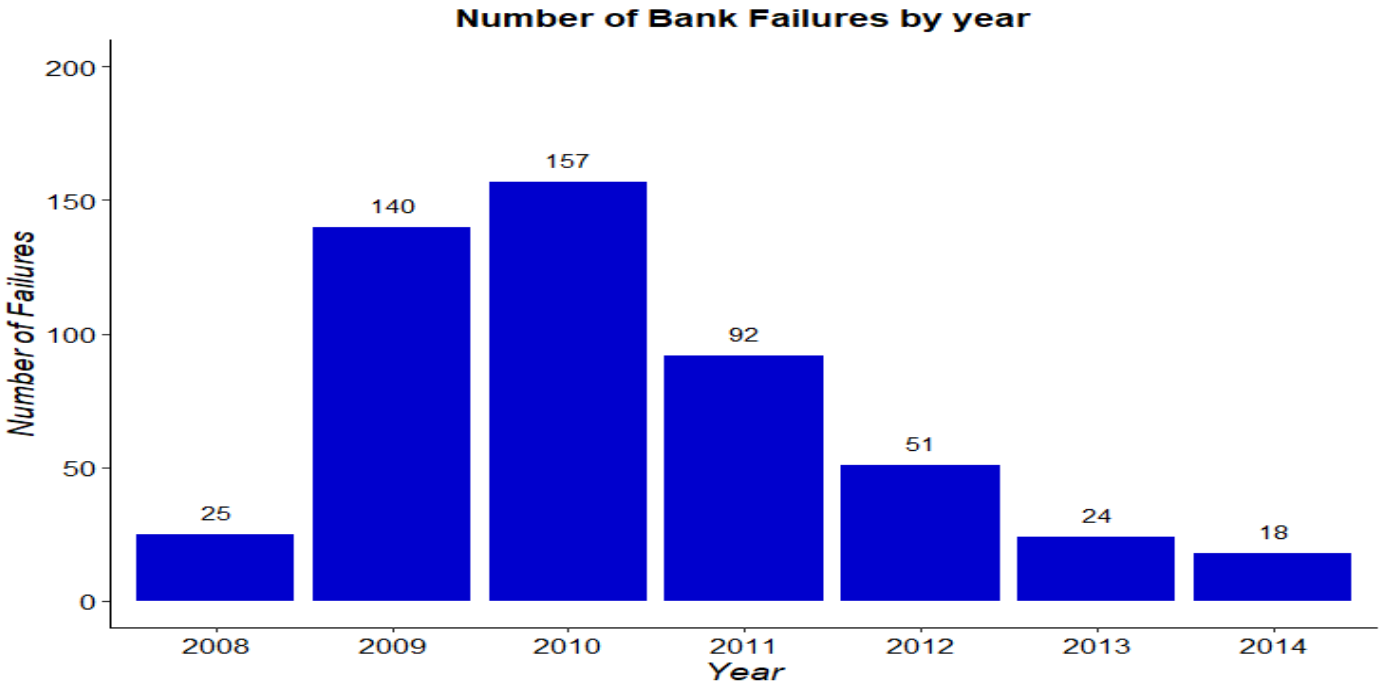

Figure 1: Bank Failures by year. 2008 - 2014

There were 9,307 N/A values present in the data, all associated with the solvent class. Due to the large nature of the data set I 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 was small. Consequently, preservation of minority class observations was the primary objective. 5,109

observations were removed in the process of resolving the outliers. After the cleaning process there were 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 (157 failures) which constituted 1.89% of the unique banks in 2010.

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

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 samples 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 the 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 performs 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 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 the papers listed as well as other papers to create a robust model. These variables are representative of the major areas of each category measured with the only CAMELS category missing was Sensitivity.

Listed in the table below is the summary statistics of the variables used in the model. As can be seen in the table there is still quite range in the variables in particular with the Efficiency Ratio and Net Loans and Leases to Core Deposits.

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
rbc1aaj	215004	10.584	3.737	-7.426	8.374	11.646	39.949
rbcrwaj	215004	17.342	7.643	-13.52	12.467	19.448	64.992
eqv	215004	11.016	3.761	-3.604	8.682	12.3	49.653
lnlsntv	215004	63.972	15.38	0.001	54.786	75.192	99.301
roe	215004	6.268	13.342	-818.735	3.097	12.004	61.542
roa	215004	0.653	1.191	-29.066	0.341	1.232	10.564
noijy	215004	0.632	1.186	-12.994	0.316	1.215	5.958
nimy	215004	3.864	0.856	-1.73	3.349	4.339	8.999
ntlnlsr	215004	0.427	0.88	-3.932	0	0.466	8.983
noniiay	215004	0.733	0.965	-3.976	0.324	0.881	19.982
eeffr	215004	73.807	33.674	-939.976	59.973	81.25	2582.927
lnatresr	215004	1.516	0.885	0	1.013	1.776	13.996
nperfv	215004	1.767	2.479	0	0.294	2.212	36.179
nclnlsr	215004	1.927	2.614	0	0.311	2.489	30.909
idlncorr	215004	96.779	47.109	0.001	73.547	113.438	1992.316
intexpy	215004	1.96	1.005	-0.002	1.107	2.75	6.497

4. Methodology

To fully understand which sampling method may prove to be the optimal choice four different sampling methods were applied representing over and under sampling. A seed of 111 was set for all sampling methods to enable reproducibility.

1.Random Under Sampling:

The process in which the majority class is reduced at random to achieve a desired ratio with the minority class. Usually resulting in a 1:1 ratio between the classes. Consider for every a_i ($i = 1 \ldots n$) observation in the minority class randomly pick b_i from the majority class until $a_i = b_i$. The training data had 372 observations from the solvent class with 372 observations from the insolvent class creating a 50:50 split.

2.Random Over Sampling with Replacement:

This is the inverse of random under sampling but altered for slightly. The minority class is sampled with replacement until the minority class achieves the desired balance usually 1:1. For every a_i ($i = 1 \ldots n$) observation in the minority class pick at random a_i ($i = 1 \ldots n$) and add it to the minority data set. Return a_i ($i = 1 \ldots n$) into the sample allowing the ability to be re-picked. Do this until the sample size is $a_i = b_i$ with b representing the majority class. This sample had 126258 insolvent and 126258 solvent observations with a 1:1 ratio.

3.Random Over Sampling Examples (ROSE):

ROSE increases the minority class by creating synthetic observations while reducing the majority class to achieve approximately a 1:1 data set. An observation is picked from the training data from either class in which the sample is created by the kernel density estimate and smoothing matrix. The choice of the kernel and smoothing matrix can be selected based on the need of research. I decided to use the default settings for the kernel and smoothing matrix.

4.Synthetic Minority Oversampling Technique (SMOTE):

An observation is picked at random from the minority class. Using K nearest neighbors, the new synthetic observation is created in the feature space created by K neighbors and placed on a line segment connecting. K is usually defaulted to 5 and has been kept as such. SMOTE has two other tuning parameters that I did attempt to tune to find the optimal balance within the data. Percent Over (OS) dictates how many extra synthetic samples should be created from the minority class. Percent Under (US) tunes how many extra observations are selected from the majority class in response to the creation of the synthetic observations from the minority class. Tuning parameters of 100% and 200% percent over sampled were tested. The arguments for percent under sampled were 25%, 50%, 75%, 100%, 150%, and 200%. The main difference between the ROSE and SMOTE sampling methods is how the synthetic samples are created. ROSE uses a smoothed bootstrapping technique while SMOTE uses KNN to place a point on the line connecting neighbors.

Logistic Regression Equation:

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

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 also known as a sigmoid function is then applied which restricts values between 0-1 and creates a linear decision boundary unless additional extensions are added to the model. Logistic Regression predicts the probability of an event transpiring. Depending on where the probability cutoff is placed, it will determine the classification of the observation. If the

probability cut off is 0.5 any value above 0.5, it will be classified as a 1 while anything below would be a 0.

Lasso Regularization:

$$\text{SSE}_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j|.$$

The Lasso penalty is commonly known as the L1 penalty. It adds a penalty of the absolute value of the magnitude of coefficients to the sum of squares error which reduces the parameter estimates. Lasso can also be used as a feature selection technique due to the ability to penalize estimates reducing coefficients to 0.

Ridge Regularization:

$$\text{SSE}_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2.$$

The Ridge Regularization is commonly known as the L2 penalty. This penalty is the square of the magnitude of the coefficients. While the L2 penalty does reduce coefficients towards 0 like L1, it will never shrink it to 0. Ridge Regularization helps to create a more robust model by preventing overfitting.

Elastic Net Regularization:

$$\mathrm{SSE}_{\mathrm{Enet}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|.$$

Elastic Net Regularization will be applied to the Logistic Regression which is a combination of the L1 and L2 penalties. The purpose of elastic net regularization is to reduce the chance of the model overfitting the training data while also dealing with any multicollinearity. Elastic Net is tuned by the mix ratio between the two with the optimal ratio being determined by the ROC_AUC for this research.

5. Validation & Results

In an attempt to find the optimal sampling method to forecast bank failures I decided upon 4 metrics in which to measure the predictive ability of the models. These metrics will help to identify the most robust method. These metrics will be used to compare the performance in the training, testing, and out of time validation set.

Validation Metrics:

Sensitivity (Recall):

Sensitivity measures how precisely the model is able to predict the positive class. This measure is of particular importance in this study due to the costs associated with potentially missing a failing bank. True Positives represent the number of positive cases (Insolvent banks)

which were predicted correctly. False Negatives represent the number of positive cases that were misclassified.

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Specificity:

Specificity calculates how accurately the true negatives are predicted correctly as compared to those who the algorithm misclassified. This will help to inform as to how well the model predicts solvent banks. True Negatives represent the number of the negative (Solvent Banks) class that were correctly classified. False Positives are the number of negative observations which were misclassified as positive.

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Balanced Accuracy:

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

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

With an imbalanced data set this imbalance in class sizes will cause accuracy to become artificially higher and not represent true predictive ability on the minority class. This is due to the

much larger portion of the equation in true negatives and false positives which will inflate the Accuracy metric. Balanced Accuracy takes into account the class distribution of the samples by taking the average accuracy by class.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

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.

Results:

Using a Penalized Logistic Regression, I analyzed the effect of four sampling methods using 16 financial ratios in the CAMELS rating systems. 12 different iterations of the SMOTE sampling method were applied in pursuit of optimization of the model. 10-Fold cross validation

was applied for robustness of the model and error terms due to James et Al. (2014) finding that k-fold validation produces more stable results than leave one out cross validation.

SMOTE Tuning Selection:

To determine which SMOTE Iterations required closer examinations they were judged based upon their performance on the Out of Time sample. The Out of Time set is the most realistic implementation of the model's performance on future samples. The main class of interest when determining the best model is the ability to correctly classify the insolvent class while minimizing false negatives. Concern does need to be given though that the model is not so focused on the minority class that it begins to forecast a large number of false positives which would decrease the resources of supervisory authorities. Depicted below are the metrics for the iterations of SMOTE on the Out of Time Sample. Focusing on sensitivity which measures how accurately the positive class is predicted, and balanced accuracy which will represent how well the model does on predicting the positive and negative class. SMOTE 200% iterations does rather well at predicting the negative class but has lackluster performance on the positive class. When comparing the two iterations of over sampling 100% over all have a better-balanced accuracy and sensitivity score. This means they are better at identifying the positive class and predict better on both classes.

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	0.9290	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

The difference in AUC for 100% over sample between the different tuning parameters of SMOTE are all rather close with the smallest only being a .0097 different from the largest. The similar case is true with balanced accuracy so the main focus then shifts to the sensitivity. There is a drop in sensitivity after under sample 75 of 0.0476 with a difference in specificity of .0091. As the minority class is of utmost importance due to the cost of failing to identify an insolvent bank is magnitudes larger than the cost of reviewing a bank the tradeoff between a higher sensitivity than specificity is a suitable compromise. It was decided the over-sample 100% with 25% under sample, 50% under sample, and 75% under sample warranted a further look and compared them to the other three sampling methods.

Comparison of Final Sampling Methods:

When exploring the results of the training sets it is clear that the SMOTE sampling methods focused on being able to predict the positive class at the expense of the negative class. This makes sense due to how the class imbalance was flipped under the SMOTE parameters. This caused the minority class to become the majority class thus attaining the class bias the solvent banks had in the regular imbalanced data set. SMOTE US 25 and 50 both have below 0.60 specificity on the solvent bank class. SMOTE US 75 has slightly less Sensitivity of 0.9530 but has a much higher Specificity along with a much higher Balanced Accuracy at 0.8474. The SMOTE iterations have a higher AUC as compared to the three other sampling methods but that can be attributed to the class bias on the insolvent class. Between the other three sampling methods Over Sampling with Replacement seems to give the most balanced result with a better predictive ability on the insolvent class. SMOTE 100 OS 75 US and Over Sampling with replacement are the top two predictive models on the testing set.

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	0.8002	0.7987	0.7995	0.8799
SMOTE 100 OS 25 US	0.9906	0.5054	0.7480	0.9467
SMOTE 100 OS 50 US	0.9718	0.5914	0.7816	0.9438
SMOTE 100 OS 75 US	0.9530	0.7419	0.8474	0.9451

The test set shows a slight degradation of predictive ability between all the sampling iterations as is expected. The ROSE sampling method though seems to show the most consistency in predictive ability with its balanced accuracy and AUC actually increasing. It became better at predicting the negative class. Random under Sampling and Over Sampling with Replacement both have a large drop in predictive ability of the positive class in the testing class. Random Under Sampling, Over Sampling with Replacement, and ROSE all have quite dismal Sensitivity but the SMOTE iterations are quite unable to correctly predict the negative class. This is quite worrying as it is important to be able to predict insolvent banks but a fine balance has to be struck.

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	0.8844
SMOTE 100 OS 25 US	0.9677	0.4506	0.7092	0.8825
SMOTE 100 OS 50 US	0.9462	0.5827	0.7645	0.8817
SMOTE 100 OS 75 US	0.9032	0.6957	0.7995	0.8844

The final test for these sampling methods is their ability to forecast on the Out of Time data set. This data set contains samples from Q1 2013 – Q4 2014 which the model has not been trained on any data points from this time period. This test seems to show a decreasing ability of Random Under Sampling, Over Sampling with Replacement, and ROSE in predicting the positive class. The balanced accuracy of all three are elevated but this is due to the model doing very well on the negative class. The three SMOTE iterations on the other hand, predict the positive and negative class rather well. SMOTE US 25 has the best sensitivity 0.8571 but it does not predict the negative classes as well. Along with that when looking back at the training and testing set it seemed the most influenced by the class bias which might not make it the most robust option. SMOTE US 50 likewise seems to be greatly affected by class bias, not so much as US 50 but based on the training and testing data it also may not be a robust enough sampling method especially if it is applied to different data sets. SMOTE US 75 has 3rd best balanced accuracy out of the group next to Over Sampling with Replacement 0.7988 and Random Under

Sampling 0.7964. SMOTE US 75 does have a leg up though as their AUC is on par but SMOTE is better at predicting the positive class which is very important. The best sampling method that could be applied to the Penalized Logistic Regression would be SMOTE 100 OS 75 US which has superior predictive ability on the Out of Time sample and decent performance on the testing set. I believe this would be the most robust model to use for forecasting bank failures.

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	0.9356
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

Discussion:

I used a rather simple Penalized Logistic Regression to test and find the optimal sampling method. Random Under Sampling, Random Over Sampling with Replacement, Random Over Sampling Example, and 12 iterations of SMOTE were tested on three data sets. By testing the model on a training, testing, and then conducting an Out of Time validation set it allowed the robustness and potential of the sampling methods to be fully explored. The Out of Time

validation set was the most important test as it was a realistic example of new quarterly data being tested on the model. Due to the parameter settings of the three SMOTE iterations compared to the rest of the sampling methods the class bias was actually reversed in favor of predicting the positive class. This led SMOTE OS 100 US 25 & SMOTE OS 100 US 50 to too severely misclassify the negative class. SMOTE OS 100 US 75 though was able to find a suitable balance between predicting the positive class and negative class. Since the cost of a bank failing is of a much larger impact than increasing supervision of solvent institutions the author believes that the degree of trade off is justified to preserve financial stability.

Limitations & Future Research:

The major limitation in this paper was personal computational issues with potential models. I originally intended to use Random Forest and Artificial Neural Networks to compare the effect of sampling methods across various classification algorithms. There is still room leverage future research on bank failures. An interesting thing would be to examine how the sampling methods perform across various classification algorithms. Along with that the US banking data sets is one of the most imbalanced sets in the world. Would the sampling methods have the same effect on a less imbalanced population in particular in regards to Under Sampling and Over Sampling that creates synthetic observations like SMOTE?

Conclusion:

The GFC renewed great interest in creating early warning systems for the financial system. Bank failures not only create reverberations in the financial markets but also in the communities they are in. This is especially true for smaller regional banks where a bank failing unexpectedly can have serious consequences on the local economy and development. The

question is will the forecasting models that are currently being worked on will be able to predict the next financial crisis? Every financial crisis in history has had a different driving force. The author believes that sampling methods could potentially be a very good method as to create manageable computational models. Work will need to be done to find a way to minimize the false positives introduced by reversing the class bias caused by the SMOTE tuning implemented in this paper as to further make the most efficient use of supervisory authorities' resources.

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

G. Karatas, O. Demir and O. K. Sahingoz, (2020)"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.

James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications. Springer, Berlin (2013)

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

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

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.

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

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

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

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.

Zhou (2013). Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods, Knowledge-Based Systems, Volume 41, Pages 16-25, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2012.12.007.

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.

Appendix

Under Sampling Confusion Matrices:

Under Sampling Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	316	67
Solvent	56	305

Under Sampling Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	70	5046
Solvent	2	26518

Under Sampling Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	31	1789
Solvent	11	54886

Over Sampling with Replacement Confusion Matrices:

Over Sampling with Replacement Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	108686	20879
Solvent	17572	105379

Over Sampling with Replacement Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	71	5237
Solvent	22	26327

Over Sample with Replacement Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	30	1878
Solvent	12	54797

ROSE Confusion Matrices:

Rose Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	50591	12761
Solvent	12633	50645

Rose Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	73	5649
Solvent	20	25915

Rose Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	30	2046
Solvent	12	54629

SMOTE $N_0 = 200$ Confusion Matrices:

SMOTE $N_0 = 200$, $N_u = 25$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	1089	87
Solvent	27	99

SMOTE $N_0 = 200$, $N_u = 25$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	90	13784
Solvent	3	17780

SMOTE $N_0 = 200$, $N_u = 25$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	31	5331
Solvent	11	51344

SMOTE $N_0 = 200$, $N_u = 50$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	1070	150
Solvent	46	222

SMOTE $N_0 = 200$, $N_u = 50$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	85	11829
Solvent	8	19735

SMOTE $N_0 = 200$, $N_u = 50$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	32	3573
Solvent	10	53102

SMOTE $N_0 = 200$, $N_u = 75$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	1048	154
Solvent	68	404

SMOTE $N_0 = 200$, $N_u = 75$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	81	7695
Solvent	12	23869

SMOTE $N_0 = 200$, $N_u = 75$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	31	2773
Solvent	11	53902

SMOTE $N_0 = 200$, $N_u = 100$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	1020	174
Solvent	96	570

SMOTE $N_0 = 200$, $N_u = 100$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	78	6534
Solvent	15	25030

SMOTE $N_0 = 200$, $N_u = 100$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	30	2266
Solvent	12	54409

SMOTE $N_0 = 200$, $N_u = 150$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	952	190
Solvent	164	926

SMOTE $N_0 = 200$, $N_u = 150$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	73	4907
Solvent	20	26657

SMOTE $N_0 = 200$, $N_u = 150$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	28	1740
Solvent	14	54935

SMOTE $N_0 = 200$, $N_u = 200$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	906	195
Solvent	210	1293

SMOTE $N_0 = 200$, $N_u = 200$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	67	3931
Solvent	26	27633

SMOTE $N_0 = 200$, $N_u = 200$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	27	1444
Solvent	15	55231

SMOTE $N_0 = 100$ Confusion Matrices:

SMOTE $N_0 = 100$, $N_u = 25$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	737	46
Solvent	7	47

SMOTE $N_0 = 100$, $N_u = 25$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	90	17342
Solvent	3	14222

SMOTE $N_0 = 100$, $N_u = 25$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	36	5872
Solvent	6	50803

SMOTE $N_0 = 100$, $N_u = 50$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	723	76
Solvent	21	110

SMOTE $N_0 = 100$, $N_u = 50$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	88	13171
Solvent	5	18393

SMOTE $N_0 = 100$, $N_u = 50$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	35	4022
Solvent	7	52653

SMOTE $N_0 = 100$, $N_u = 75$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	709	72
Solvent	35	207

SMOTE $N_0 = 100$, $N_u = 75$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	84	9604
Solvent	9	21960

SMOTE $N_0 = 100$, $N_u = 75$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	34	3299
Solvent	8	53376

SMOTE $N_0 = 100$, $N_u = 100$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	701	43
Solvent	76	296

SMOTE $N_0 = 100$, $N_u = 100$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	80	8073
Solvent	13	23491

SMOTE $N_0 = 100$, $N_u = 100$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	32	2783
Solvent	10	53892

SMOTE $N_0 = 100$, $N_u = 150$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	676	99
Solvent	68	459

SMOTE $N_0 = 100$, $N_u = 150$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	76	6436
Solvent	17	25128

SMOTE $N_0 = 100$, $N_u = 150$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	32	2359
Solvent	10	54316

SMOTE $N_0 = 100$, $N_u = 200$ Training:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	648	118
Solvent	96	626

SMOTE $N_0 = 100$, $N_u = 200$ Testing:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	71	5166
Solvent	22	26398

SMOTE $N_0 = 100$, $N_u = 200$ Out of Time:

Prediction	Reference	
	Insolvent	Solvent
Insolvent	32	1922
Solvent	10	54753