Predicting Bank Failure:

Final Report for the CSBS Data Analytics Competition

Ignacio, Romeo[1], Xie, Yunjie[2], and Devarajan, Abhishek[3]

[1]University of California, Irvine

[2]University of California, Irvine; Renmin University of China

[3]University of California, Irvine

November 22, 2021

Advisor: Professor Gary Richardson

Abstract

The health and efficiency of the US economy is strongly associated with the economic health of banks. In order to avoid another financial crisis such as the 2008 housing market crash, it is crucial to monitor the banking sector and spot potential bank failures before they occur. Typically, this prediction process is very difficult, due to (1) the vast number of metrics associated with each bank and (2) the very low rate of failure among banks. In this paper, we use risk scoping data from the Conference of State Bank Supervisors (CSBS), as well as a variety of other bank-related data, to build a statistical model that predicts bank failures. Using a CAMELS-like method of classifying banking metrics, we find that during times of distress, such as the 2008 crisis, our model can accurately predict failure. Earnings variables, in particular, are strong predictors of a bank's outcomes. Outside of periods of distress, however, we find that predicting bank failure is difficult. Furthermore, we find that risk levels are not good predictors of bank failure more than two quarters in the future. The amount of risk that a bank currently has is not a good predictor of the amount of risk a bank will have more than two quarters into the future.

1 Introduction

During the 2008 financial crisis, an abnormally high number of banks failed in the US, which froze credit markets and resulted in massive slowdowns in both the domestic and international economy (Erkens et al., 2012). The bank failures of this time period greatly contributed to the Great Recession, which was the second largest recession in US history, after the Great Depression. In 2020, as the COVID-19 pandemic causes the US economy to suddenly slow down, the ability of banks to stay in business is of vital importance. In order to avoid another banking crisis and another huge recession, it is crucial to detect potential bank failures before they occur, and to take the necessary precautions to avoid them.

With that being said, the following question remains. Is it possible to predict bank failure based on prior data and, if so, what data do we need to make such predictions? This is the question that our team set out to answer in this paper. In order to find a suitable answer, we utilize data from the Conference of State Bank Supervisors (CSBS), the Federal Deposit Insurance Corporation (FDIC) and the Federal Reserve Economic Data (FRED). With these datasets, we need to overcome two main challenges. First, there is the fact that very few banks fail, even in times of crisis. Those that do fail do so sporadically, in such a way where overarching patterns and links between multiple banks failing are difficult to find. Second, we must find a way to take the large number of variables present in the datasets and narrow them down to a small set of variables to use for our predictive model. This is challenging because many of the variables are highly correlated with one another, which leads to multicolinearity. Often, this multicolinearity results from the fact that balance sheet variables are simply linear combinations of other variables. Adding multiple variables that are closely correlated with each other is likely to decrease the accuracy and robustness of our model. In addition, the vast number of variables to select from presents a risk of having an over-identified dependent variable. In other words, there are too many possible explanations for why a bank failed.

To overcome these hurdles, we focus on predicting failures during the financial crisis of 2008 and the three years after, i.e. the period from 2008Q2 up to 2011Q2. In this period, many banks failed, which helps us to overcome problems posed by low bank failures rates in typical time periods. We believe that this in-sample period contains the most relevant information in regards to predicting when a bank will fail, as well as which factors contributed to said failure. We use

Principal Components Analysis (PCA) to reduce the dimensionality of our dataset while preserving as much information as possible. Furthermore, we use a t-test for variable selection, further refining the number of variables within our model.

The model we use for this project is a logistic regression (logit) model in Stata. We build four variations of the logit model in order to test various combinations of independent variables. Model 1 is a Discrete Time Hazard Model that uses Principal Components (PCs) of call report variables, as well as a time-fixed effect in order to control for factors that influence the baseline hazard rate. Model 2 also uses the PCs of call report variables but substitutes the fixed-time effect with PCs of macroeconomic time series variables. Model 3 is another Discrete Time Hazard Model, but instead of using Principal Components, we use variables from the Risk Scoping dataset. Similarly, Model 4 is analogous to Model 2, but uses Risk Scoping variables in place of PCs. We fit each model using the in-sample data, and then make predictions. Finally, we test the validity of these predicitions by using a t-test to look for Type I and Type II error.

We find that Models 1 and 2 do an excellent job of predicting failure both in and out of sample; however, these estimates do not yield specific directions that regulators can examine. Models using Risk Scoping variables have decent predictive capabilities, but are not quite as good as the PCA models. Models 3 and 4 do have an advantage in terms of interpretability, as they are comprised of specific items on bank balance sheets which predict failure. In our case, we find that the most significant variables for predicting failure relate to earnings. Banks that have the lowest earnings are the most likely to fail.

All four of our models have difficulty predicting failure outside of crisis times, when failure rates are very low. An extreme example is that many quarters have no failures, but our model always predicts that the worst banks will fail. As such, our model often overpredicts failure in normal times. This is an important issue and, as such, we explore what factors limit the accuracy of our model during normal times.

We do this by creating a variable that tracks the total number of exceptions– or red flags– for each bank at each quarter. We then construct another logit model using the number of exceptions as the independent variable. With this model, we perform two regressions: one to predict failure, and one to predict future exceptions. We find that exceptions are good predictors of failure for a quarter or two, but their predictive power wanes after that. Similarly, we also find that current levels of

exceptions have no correlation with future levels of exceptions after a period of two quarters. This result indicates that the preponderance of banks that are flagged with exceptions correct those exceptions. The few that do not usually merge. Only a small percentage of banks that take neither of those two options end up failing.

The rest of our essay proceeds as follows. Section 2 describes the datasets we used for analysis. Section 3 outlines our methods of consolidating our data using Principal Component Analysis . Section 4 discusses the four models we established to predict banking failure. Section 5 shows how well these models correctly and incorrectly assigned failed and non failed banks. Lastly, Section 6 gives a general discussion of our paper and future avenues of research.

2 Data

For our analysis, we utilize a collection of four different datasets. These datasets were sourced from various organizations (listed in Section 1) and served different purposes in our analysis. In addition to collecting the raw data, we performed a number of data cleaning and transformation processes to better tailor the data to our model. More information about each dataset can be found below, in the following subsections.

2.1 Risk Scoping

The primary dataset that we use for this project is the Risk Scoping dataset provided by CSBS. This dataset consists of 9316 banks and provides a wide selection of balance sheet information for each bank. In total, there are 112 different variables associated with each bank. The risk scoping dataset follows a time-series format, with entries for each bank beginning in 2006Q1 and ending in 2020Q3. In the case of a bank failure or merger, the bank in question is removed from the dataset starting in the following quarter. That is, if Bank X failed in 2012Q2, then from 2012Q3 onward, there will be no information reported for Bank X.

In addition to the raw dataset, the CSBS provides us an excel workbook that explains how they utilize the risk scoping data to monitor financial institutions. The workbook includes formulas used to ratios from variables in the risk scoping set. There are also thresholds for each variable that allow us to determine whether a bank is assuming a low, moderate or high amount of risk within

that variable. Using the formulas from the excel workbook, we calculate the ratios listed and add them as new columns to the risk scoping dataset. Then, with the help of the thresholds in the workbook, we calculate a second set of variables based on the variables provided in the risk scoping set. This secondary set of variables, labeled with the prefix "thresh", essentially collapses the wide ranges of values within each column to one of three possible values:

- 1 - Red Flag (High Risk)
- 2 - Yellow Flag (Moderate Risk)
- 3 - Green Flag (Low Risk)

From here, we create another column, TotalExceptions, that approximates the total risk level for each bank by using the risk levels associated with each of the banks' variables. We do this by summing up the number of red flags associated with each bank. We expect that this total number of exceptions is negatively correlated with the likelihood of bank failure, as a lower value for TotalExceptions indicates that a bank is taking on a large amount of risk.

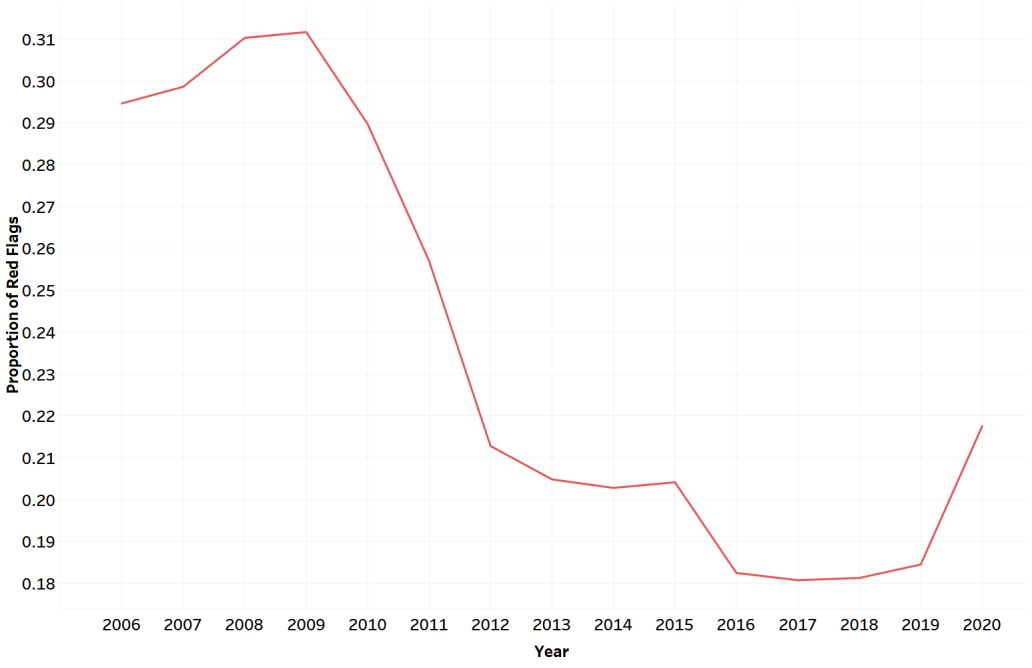

Figure 1: Proportion of Red Flags by Year

In the figure above, we depict the number of exceptions, or red flags, that occur within each

year. As one would expect, we see a much higher number of red flags during the crisis time period compared to the subsequent time period.

2.2 FDIC Failure Data

Although the Risk Scoping dataset provides a large number of variables that are useful for predicting failure, the set lacks a proper dependent variable. To remedy this, we use Failure Data from the FDIC to construct an indicator variable named Failed. This variable remains 0 for all quarters until the bank fails. During the quarter at which the bank fails, the variable takes the value 1.

Table 1: Banking Outcome by Year

	Outcome						
	Survived		Merged		Failed		Total
	No.	Percent	No.	Percent	No.	Percent	No.
2006	34,700	99.04%	333	0.95%	2	0.01%	35,035
2007	34,060	99.02%	332	0.97%	5	0.01%	34,397
2008	33,349	99.04%	280	0.83%	42	0.12%	33,671
2009	32,255	98.98%	192	0.59%	142	0.44%	32,589
2010	30,850	98.82%	243	0.78%	126	0.40%	31,219
2011	29,649	99.10%	188	0.63%	80	0.27%	29,917
2012	28,564	99.00%	254	0.88%	35	0.12%	28,853
2013	27,407	98.95%	268	0.97%	23	0.08%	27,698
2014	26,209	98.83%	292	1.10%	19	0.07%	26,520
2015	24,957	98.82%	293	1.16%	5	0.02%	25,255
2016	23,841	98.89%	261	1.08%	7	0.03%	24,109
2017	22,831	98.89%	250	1.08%	6	0.03%	23,087
2018	21,817	98.87%	250	1.13%	0	0.00%	22,067
2019	20,876	98.77%	255	1.21%	5	0.02%	21,136
2020	15,146	99.37%	85	0.56%	11	0.07%	15,242
Total	406,511	98.96%	3,776	0.92%	508	0.12%	410,795

Data Source: FDIC-BankFind Suite: Bank Failures & Assistance Data

After acquiring the failure data, we were able to examine the distribution of failures across time. The table above highlights the percent of banks that failed and survived for each year. It is clear that for every quarter, the number of failures is minuscule. This holds even for years during the financial crisis. The lack of failures in the dataset creates an issue in terms of model development. Specifically, if we attempt to build a model that focuses on having the highest

accuracy of predictions, then it is very likely that our model will simply predict "not failed" for every single bank. If this were the case, only around 0.12% of predictions would be incorrect, yet our model would not tell us anything of value.

2.3 FDIC Call Reports

Our third dataset is a much larger set of banks and various balance sheet data. Once again, this info comes from the FDIC and it spans the time period from 1993-2020. We use the variables from this dataset as a test to see if the Risk Scoping variables are truly the best choice to look at for predicting failure.

2.4 Macroeconomic Data

Finally, we incorporate a set of macroeconomic variables from the Federal Reserve into our project. We used the database from McCraken and Ng (2020) as they established a rich set of macroeconomic variables for our period of interest. The macroeconomic variables are mainly used to look for correlations or explanations for various trends or anomalies in our models. This will be discussed in further detail within our Model and Results sections. The variables we use from this dataset include GDP and the national unemployment rate.

3 Methods

3.1 Principal Components

As the introduction stated, one of the fundamental issues that we face in this study is the fact that we have an enormous amount of data on each bank. Consequently, if we use all of the variables available to us, failure may be over-identified. By this, we mean that there may be many variables that are good predictors of failures, but are strongly correlated with each other. This strong correlation can be explained by two factors. The first contributing factor is that many of the call report variables are representative of underlying patterns of choices made by bankers. Secondly, the correlation is due to the fact that many variables are simply linear combinations or ratios of other variables; thus, they are directly dependent on one another. Adding all of these variables to our model can lead to multicolinearity and a higher rate of false-negative predictions.

Furthermore, arbitrarily or hueristically choosing variables is not likely to avoid multicolinearity entirely, and may leave out relevant statistical information. These complications raise the question of what technique can be used to reduce the number of correlated variables, while maintaining as much information as possible. Our answer to these questions are that circumventing is Principal Component Analysis (PCA).

PCA is a technique by which a large dataset with a high number of dimensions is "reduced" to a smaller set of variables, while preserving as much of the original dataset's variance as possible. Sources such as Applied (2021a) from Penn State, Applied (2021b) from Penn State, and Shlens (2021) explain what and how PCs are in a more detailed manner and were sources that we consulted when understanding the theory behind a PCA. The idea of a PCA is to collapse all data into independent eigenvectors using the matrix transformation as follows:

$$\mathbf{Y} = \mathbf{XP}$$

where $\mathbf{X}$ is our original data matrix with dimensions n by m, $\mathbf{P}$ is an m by m orthogonal matrix consisting of eigenvectors of $\mathbf{X}$, and $\mathbf{Y}$ is our newly transformed data with dimensions n by m. Matrix $\mathbf{Y}$ can be thought of as a weighted average of vectors in matrix $\mathbf{X}$. In our case, n is the number of quarterly banking observations, m is the number of banking variables we had in our set.

We will describe the creation of PCs in terms of Singular Value Decomposition (SVD) as the article Applied (2021a) and Shelns (2021) stated. SVD is a way to decompose a matrix into eigenvectors and eigenvalues as follows:

$$\mathbf{X} = \mathbf{UDP}^T$$

where $\mathbf{U}$ is an orthogonal matrix with dimensions n by n, and $\mathbf{D}$ is diagonal matrix with eigenvalues such that the diagonal elements are organized from highest to lowest value with dimensions n by p.

The sample variance covariance from Applied (2021a) is as follows:

$$\mathbf{Var}(\mathbf{X}) = \mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{N}$$

When we plug in the SVD of $\mathbf{X}$ we get:

$$\mathbf{S} = \mathbf{P}\mathbf{D}^2\mathbf{P}^T$$

Here, $\mathbf{D}^2$ is simply the matrix $\mathbf{D}$ multiplied by itself. Since $\mathbf{D}$ is a diagonal matrix, this means that $\mathbf{D}^2$ is also diagonal, with each entry being the square of its corresponding entry in $\mathbf{D}$.

From (1) , (2), Applied (2021a), and Shelns (2021), the variance of the eigenvector matrix $\mathbf{V}$ is:

$$\begin{aligned}
&\mathbf{Var}(\mathbf{Y}) \\
=&\mathbf{Var}(\mathbf{XP}) \\
=&\mathbf{P}^T\mathbf{Var}(\mathbf{X})\mathbf{P} \\
=&\mathbf{P}^T\mathbf{X}^T\mathbf{XP} \\
=&\mathbf{P}^T\mathbf{P}\mathbf{D}^2\mathbf{P}^T\mathbf{P} \\
=&\mathbf{D}^2
\end{aligned}$$

Since $\mathbf{P}$ is an orthogonal matrix, $\mathbf{P}^T\mathbf{P} = I$ (this is a theorem found in Shelns (2021)). What's interesting to note is that the set of eigenvalues that solves this matrix is unique, but this is not necessarily true of the eigenvector; any eigenvector can be scaled by an arbitrary factor, and it will remain an eigenvector. We also organize our variance covariance matrix in such a way where the first eigenvector in $\mathbf{P}$, which corresponds with the first eigenvalue in matrix $\mathbf{D}^2$, explains the most variance. Similarly, the next eigenvector explains the second most variance of the data, and so on. Finally, we would like to note that eigenvectors are linearly independent of each other and, as such, form a new basis for the matrix.

For our FDIC created set with 200 variables, we standardize all variables with mean zero and variance 1 using z-scores. We then estimate the PCs for these standardized variables. For our macroeconomic set from McCraken and Ng (2020) with 100 variables, we predict PCs with standardized and non standardized variables macroeconomic variables. We will only use a subset of these PCs to estimate with the first PC estimated explaining the most variation in the set, the second PC explaining the second most variation in the set, and so on. To determine the amount

of PCs to estimate, we used the eigenvalue of 1 as a cutoff. For our FDIC set, we collapsed the original 200 variables into 24 PCs. For the macroeconomic set, we collapsed these 100 variables into 10 PCs. We found that our models had better predictions when macroeconomic variables PCs were predicted with non standardized variables. We are unsure why this occurred and we encourage any researcher to look into these differences between standardized and non standardized macroeconomic variables PC predictions.

Table 2: Rotated PCA

PCA1	Net Loans, Deposits, Good Will, Assets	PCA13	Performance
PCA2	Assets	PCA14	Loan Charges, Past Due Loans
PCA3	Income	PCA15	Total Deposits
PCA4	Past Due (Loans), Loan Charges	PCA16	Assets, Performance
PCA5	Loan Charges	PCA17	Assets, Net Loans, Performance
PCA6	Securities	PCA18	Performance
PCA7	Loan Charges, Net Loans	PCA19	Loan Charges
PCA8	Additional Noninterest Income, Loan Charges	PCA20	Loan Charges, Performance
PCA9	Income, Net Loans, Past Due Loans	PCA21	Performance
PCA10	Assets and Liabilities, Real Estate, Past Due Loans	PCA22	Loan Charges
PCA11	Net Loans, Securities	PCA23	Additional
PCA12	Past Due Loans, Good Will	PCA24	Loan Charges

As mentioned previously, one of the main problems with PCA is the fact that it is difficult to interpret what each principle component means. Even though each principal component is merely a linear combination of the variables within our dataset, the weights assigned to each variable are difficult to ascribe any significance to. A solution to this problem is to "rotate" the principal components. Using the same change of basis mathematics that PCA used, this procedure will modify the weights assigned to each variable to make them easier to interpret. After rotating the components, we look at the weight associated with each variable and select the variables with the largest, most positive weights. These variables serve as the interpretation of our principle components. In essence, we assume that the most highly weighted variables are the driving factors behind the model's ability to predict bank failures.

3.2 Variable Selection

PCAs: We choose raw variables based on a few criteria to determine a PC. A variable must have almost all observations for both failed and non failed banks (greater than 95% of valid observations).

It is especially important that failed bank observations are almost complete when predicting PCs as variables with a large amount of missing observations for failed banks will not be able to predict PCs of failed banks. We predict all PCs for any PCs with an eigenvalue greater than one.

Models 1 & 2: For Models 1 and 2, we utilize several criteria in order to determine whether to keep or take out a PC in the model. A PC that either explains a majority of variance or has a statistically significant difference based on t-tests between failed and non-failed groups is kept in the model. All other PCs are discarded.

Models 3 & 4: For Models 3 & 4, we determined whether or not to keep a risk-scoping variable if it had a statistically significant difference based on t-tests between failed and non-failed groups.

4 Models

The main model we use in this study is a logistic regression model. Given that we are trying to predict bank failure, the dependent variable for the model is the indicator variable *Failed*. In order to account for external factors such as the state of the economy and the specific quarter, we will fit our model using four sets of independent variables. More specific details about each variant of our model will be available in the subsections below.

4.1 Model 1: PCs of Bank Characteristics & Time Hazard

Our first model attempts to predict the outcome of bank i at time t, which we call f_{it} using the Principal Components of bank characteristics from the FDIC call reports (denoted as P_i) as well as an indicator variable for the quarter I_t. The model follows the formula below:

$$\textbf{logit}(f_{it}) = \beta_0 + \sum_j \beta_j P_{itj} + \sum_t \alpha_t I_t + \epsilon_{it}$$

Here, ϵ_{it} represents the error associated with each bank i at time t.

As mentioned earlier, we use PCA in order to reduce the dimensionality of our dataset. This is a crucial step in making our model viable for everyday use. With over 200 variables in the FDIC call reports, a huge amount of computational power would be necessary to run the model using all variables. In addition, hand selecting variables to reduce the number of parameters in the model

would either take just as much time, or result in completely arbitrary variable section. Thus, we believe that PCA is the key to retaining the statistical information of our dataset while shrinking the size of our model.

4.2 Model 2: PCs of Bank Characteristics & Macroeconomic Variables

Our second model follows the same general form of the first model. However, instead of an indicator variable for the quarter, we add Principal Components of macroeconomic variables at time t (M_t).

$$\textbf{logit}(f_{it}) = \beta_0 + \sum_j \beta_j P_{itj} + \sum_k \beta_k M_{tk} + \epsilon_{it}$$

Table 3: Estimates of Model 1

Variables	Coefficient	Standard Error	dy/dx	Standard Error
PCA1	-0.849	0.101	-0.002	0.000
PCA2	0.406	0.080	0.001	0.000
PCA3	-0.585	0.124	-0.001	0.000
PCA5	-1.278	0.196	-0.003	0.001
PCA6	-1.212	0.080	-0.003	0.000
PCA7	-1.014	0.157	-0.003	0.000
PCA8	0.403	0.060	0.001	0.000
PCA10	0.754	0.099	0.002	0.000
PCA11	0.353	0.022	0.001	0.000
PCA13	1.712	0.132	0.004	0.000
PCA14	1.275	0.137	0.003	0.000
PCA16	-1.013	0.136	-0.003	0.000
PCA17	0.227	0.087	0.001	0.000
PCA18	-0.549	0.091	-0.001	0.000
PCA19	1.121	0.107	0.003	0.000
PCA22	0.968	0.073	0.002	0.000
PCA23	-0.167	0.067	0.000	0.000
_cons	-8.569	0.392		
Observation	103,960	Prob > chi2	0.0000	
LR chi2(33)	2,205.510	Pseudo R2	0.458	

Table 4: Estimates of Model 2

Variables	Coefficient	Standard Error	dy/dx	Standard Error
PCA1	-0.857	0.099	-0.002	0.000
PCA2	0.449	0.085	0.001	0.000
PCA3	-0.603	0.117	-0.002	0.000
PCA5	-1.120	0.198	-0.003	0.001
PCA6	-1.234	0.081	-0.003	0.000
PCA7	-1.007	0.152	-0.003	0.000
PCA8	0.383	0.060	0.001	0.000
PCA10	0.705	0.092	0.002	0.000
PCA11	0.348	0.022	0.001	0.000
PCA13	1.632	0.134	0.004	0.000
PCA14	1.429	0.156	0.004	0.000
PCA16	-0.938	0.139	-0.002	0.000
PCA17	0.204	0.086	0.001	0.000
PCA18	-0.506	0.089	-0.001	0.000
PCA19	1.164	0.105	0.003	0.000
PCA22	0.904	0.068	0.002	0.000
PCA23	-0.192	0.054	0.000	0.000
PCA24	-0.216	0.076	-0.001	0.000
_cons	-6.773	2.087		
Observation	103,960	Prob > chi2	0.0000	
LR chi2(25)	2,199.220	Pseudo R2	0.456	

In these two tables, we list the set of principal components that were deemed statistically significant without our first and second models. We find that both models have nearly the same selection of PCs, with the slight exception that Model 2 also contains PCA24, while Model 1 does not. In the columns labeled "dy/dx" we document the marginal effects of each components on the predictions that our models make.

4.3 Model 3: Risk Scoping Variables & Time Hazard

For model 3, we substitute the FDIC call reports for the risk scoping dataset. In order to provide more interpretability to our model, we choose to use the risk scoping variables as they are, rather than calculating the PCs of the risk scoping dataset. Thus, the third model uses the following equation:

$$\textbf{logit}(f_{it}) = \beta_0 + \sum_j \beta_j RS_{itj} + \sum_t \alpha_t I_t + \epsilon_{it}$$

Note that we use RS_{it} to represent the risk scoping variables for bank i at time t.

4.4 Model 4: Risk Scoping Variables & PCs of Macroeconomic Variables

Our final model is analogous to Model 2, with risk scoping variables replacing the PCs of back characteristics:

$$\textbf{logit}(f_{it}) = \beta_0 + \sum_j \beta_j RS_{itj} + \sum_k \beta_k M_{tk} + \epsilon_{it}$$

Table 5: Variables in Model 3 & 4

Variable	Label
InvScope1	Total Securities/ Assets (%)
AQScope4	Loan Loss Reserves/ Total Loans and Leases (%)
EarnScope1	Return on Average Assets (%)
EarnScope2	Net Interest Margin (%)
EarnScope5	Gain on Securities/ Avg Assets (%)
EarnScope6	Provision Expense/ Avg Assets (%)
LiqScope16	K434 Core Deposits / Total Deposits (UBPR)

In this table, we display the set of variables used to build out third and fourth models. Compared to the first two models, Models 3 and 4 are easier to interpret because their independent variables are directly picked from the Risk Scoping dataset. However, the downside to this is that these variables may not explain much of the variation within the Risk Scoping dataset.

Table 6: Estimates of Model 3

Variable	Coefficient	Standard Error	dy/dx	Standard Error
Total Securities/ Assets (%)	-0.090	0.016	-0.0003	0.0000
Loan Loss Reserves/ Total Loans and Leases (%)	0.055	0.022	0.0002	0.0001
Return on Average Assets (%)	-0.078	0.023	-0.0002	0.0001
Return on Average Assets (%) (t-1)	-0.105	0.043	-0.0003	0.0001
Net Interest Margin (%)	-0.389	0.085	-0.0011	0.0002
Net Interest Margin (%) (t-1)	-0.318	0.088	-0.0009	0.0003
Net Interest Margin (%) (t-2)	-0.204	0.075	-0.0006	0.0002
Gain on Securities/ Avg Assets (%)	-0.234	0.106	-0.0007	0.0003
Gain on Securities/ Avg Assets (%) (t-1)	-0.287	0.125	-0.0008	0.0004
Provision Expense/ Avg Assets (%)	0.332	0.042	0.0009	0.0001
Provision Expense/ Avg Assets (%) (t-2)	0.205	0.051	0.0006	0.0001
K434 Core Deposits / Total Deposits (UBPR)	0.038	0.007	0.0001	0.0000
K434 Core Deposits / Total Deposits (UBPR) (t-2)	-0.023	0.008	-0.0001	0.0000
_cons	-5.709	0.360		
Observation	103,833	Prob > chi2	0.0000	
LR chi2(42)	1,339.04	Pseudo R2	0.2866	

Table 7: Estimates of Model 4

Variable	Coefficient	Standard Error	dy/dx	Standard Error
Total Securities/ Assets (%)	-0.090	0.016	-0.0003	0.0000
Loan Loss Reserves/ Total Loans and Leases (%)	0.056	0.022	0.0002	0.0001
Return on Average Assets (%)	-0.079	0.023	-0.0002	0.0001
Total Securities/ Assets (%) (t-1)	-0.103	0.042	-0.0003	0.0001
Net Interest Margin (%)	-0.376	0.082	-0.0011	0.0002
Net Interest Margin (%) (t-1)	-0.337	0.071	-0.0010	0.0002
Net Interest Margin (%) (t-2)	-0.208	0.068	-0.0006	0.0002
Gain on Securities/ Avg Assets (%)	-0.232	0.105	-0.0007	0.0003
Gain on Securities/ Avg Assets (%) (t-1)	-0.292	0.124	-0.0008	0.0004
Provision Expense/ Avg Assets (%)	0.331	0.042	0.0009	0.0001
Provision Expense/ Avg Assets (%) (t-1)	0.118	0.060	0.0003	0.0002
Provision Expense/ Avg Assets (%) (t-2)	0.204	0.051	0.0006	0.0001
K434 Core Deposits / Total Deposits (UBPR)	0.038	0.007	0.0001	0.0000
K434 Core Deposits / Total Deposits (UBPR) (t-2)	-0.022	0.008	-0.0001	0.0000
Observation	103,833	Prob > chi2	0.0000	
LR chi2(39)	1,338.22	Pseudo R2	0.2864	

Similar to table 3 and table 4, we use these tables to showcase the statistically significant variables for our third and fourth models. Some of the variables in these tables have markers that read $(t-1)$ or $(t-2)$. These markers indicate a time lag of one or two quarters. We implement these time-lagged variables in our models in order to test whether the correlation between a variable and bank failure changes over time. Realistically, issues that banks face will take time to manifest, and they can go undetected for many quarters before ultimately leading to the failure of the bank. For our models, we limit the lag to only two quarters because of the huge computational effort needed to introduce further lag. In particular, with 3-4 quarters of lag, our models take upwards of 12

hours to run.

4.5 Exceptions Regressions

$$logit(r_{it}) = \beta_0 + \sum_t \beta_t r_{it} + \epsilon_{it} \qquad (*)$$

$$logit(f_{it}) = \beta_0 + \sum_t \beta_t r_{it} + \epsilon_{it} \qquad (**)$$

Along with the main model, we also perform a set of regressions using the number of total exceptions, r_{it} for each bank i at time t. Note that in both (*) and (**), we sum across all periods less than or equal to the period we are predicting. First, we use the model described in (*) to predict future red flags given a bank's current number of red flags. We then use model (**) to predict failure using total exceptions.

5 Results

5.1 Model Accuracy

Table 8: Accuracy of Model 1

PCA and Time Hazard Model

In-Sample 2008Q2-2011Q2

	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	15777	87941	85%	14745	88973	86%	13862	89856	87%	12653	91065	88%	7638	96080	93%
Failed	36	326	90%	55	307	85%	106	256	71%	150	212	59%	300	62	17%
Total	15813	88267	100%	14800	89280	100%	13968	90112	100%	12803	91277	100%	7938	96142	100%

Out-of-Sample 2012-2014

	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	28627	54371	66%	28259	54739	66%	26769	56229	68%	24617	58381	70%	18866	64132	77%
Failed	2	71	97%	3	70	96%	3	70	96%	4	69	95%	14	59	81%
Total	28629	54442	100%	28262	54809	100%	26772	56299	100%	24621	58450	100%	18880	64191	100%

Table 9: Accuracy for Model 2

PCA and Macro Model

In-Sample 2008Q2-2011Q2

	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	20254	83464	80%	19412	84306	81%	19178	84540	82%	19776	83942	81%	26992	76726	74%
Failed	29	333	92%	51	311	86%	94	269	74%	132	230	64%	251	111	31%
Total	20283	83797	100%	19463	84617	100%	19272	84809	100%	19908	84172	100%	27243	76837	100%

Out-of-Sample 2012-2014

	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	6774	76224	92%	7099	75899	91%	7621	75377	91%	8412	74586	90%	12055	70943	85%
Failed	3	70	96%	5	68	93%	5	68	93%	6	67	92%	15	58	79%
Total	6777	76294	100%	7104	75967	100%	7626	75445	100%	8418	74653	100%	12070	71001	100%

In this set of tables, we display the accuracy of our models across multiple time periods. For each model, accuracy is reported for the in-sample period (2008Q2 - 2011Q2) as well as the out-of-sample period (2012-2014).

Looking at Model 1, we can see that the in-sample accuracy decays rapidly. Moving from out quarter out to three quarters out drops the percentage of correct failure predictions from 90% to 71%. Going out even further than three quarters leads to another drastic drop in accuracy. However, when we look at the out-of-sample accuracy, we find a much different story. Model 1 is able to consistently predict failures with over 90% accuracy for an entire year. Even after two years, the accuracy is still around 80%.

The case for Model 2 is similar. The in-sample accuracy seems to deteriorate quickly as we attempt to predict failures further than one quarter out. Although the accuracy for each time period is slightly higher than their Model 1 counterparts, they are still unremarkable. With out-of-sample data, however, Model 2 is able to make accurate predictions up to two years out, just as Model 1 was able to do.

Along with the accuracy, we also use these tables to report the rates of Type I and Type II error. In the context of this project, we define Type I error as a False Positive (i.e. the model predicts that a bank will fail even when it doesn't) and Type II error as a False Negative (i.e. the model predicts that a bank will survive when it actually fails). We note that for the purposes of preventing bank failures, having a higher rate of Type I error is not as detrimental as having a higher rate of Type II error. In the end, it is better for regulators to be overly cautious and inspect healthy banks than it is to have failing banks go unnoticed. Looking at the prevalence of each type of error for Models 1 and 2, we see that there are relatively few Type II errors, compared to

the number of Type I errors. This holds for both in-sample and out-of-sample data, and persists regardless of how far out predictions are made. Comparing the two models, we find that Model 1 has a significantly higher rate of Type I errors than Model 2, even though both have similar rates of Type II error.

Table 10: Accuracy of Model 3

RS and Time Hazard Model

In-Sample 2008Q2-2011Q2															
	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	23914	79817	77%	22910	80821	78%	22333	81398	78%	21200	82531	80%	23617	80114	77%
Failed	30	319	91%	62	287	82%	95	254	73%	136	213	61%	251	98	28%
Total	23944	80136	100%	22972	81108	100%	22428	81652	100%	21336	82744	100%	23868	80212	100%
Out-of-Sample 2012-2014															
	1 Quarter Out			**2 Quarters Out**			**3 Quarters Out**			**1 Year Out**			**2 Years Out**		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	3327	79670	96%	4163	78834	95%	6216	76781	93%	7550	75447	91%	14785	68212	82%
Failed	41	33	45%	37	37	50%	28	46	62%	28	46	62%	23	51	69%
Total	3368	79703	100%	4200	78871	100%	6244	76827	100%	7578	75493	100%	14808	68263	100%

Table 11: Accuracy of Model 4

RS and Macro Model

In-Sample 2008Q2-2011Q2															
	1 Quarter Out			2 Quarters Out			3 Quarters Out			1 Year Out			2 Years Out		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	23169	80562	78%	21210	82521	80%	19732	83999	81%	17826	85905	83%	17851	85880	83%
Failed	36	313	90%	70	279	80%	105	244	70%	150	199	57%	282	67	19%
Total	23205	80875	100%	21280	82800	100%	19837	84243	100%	17976	86104	100%	18133	85947	100%
Out-of-Sample 2012-2014															
	1 Quarter Out			**2 Quarters Out**			**3 Quarters Out**			**1 Year Out**			**2 Years Out**		
	Incorrect	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*	*Incorrect*	*Correct*	*% Correct*
Not Failed	16456	66541	80%	16951	66046	80%	15977	67020	81%	16485	66512	80%	19874	63123	76%
Failed	5	69	93%	11	63	85%	10	64	86%	7	67	91%	13	61	82%
Total	16461	66610	100%	16962	66109	100%	15987	67084	100%	16492	66579	100%	19887	63184	100%

The accuracy tables for Models 3 and 4 are quite interesting when compared to the tables for Models 1 and 2. Looking at Model 3 first, we can see that the in-sample accuracy for predicted failures follows a similar decay pattern to the first and second models. However, when looking at the out-of-sample tests, we see a much different result. Unlike the first two models, which either steadily decreased in accuracy or maintained accuracy as predictions were made further out, Model 3 seems to have a slight increase in accuracy as predictions are made further into the future. Even with this slight upward trend, however, Model 3 is still not able to measure up to the accuracy scores of Model 1 or Model 2. At its best, Model 3 has a 69% accuracy when predicting failures that will occur two years in advance.

Model 4 returns to the same accuracy patterns that we see with Model 1 and Model 2. The out-of-sample tests outperform the in-sample tests, and the model is able to maintain accurate

predictions up to two years out when using out-of-sample data.

Looking at the Type I and Type II errors, it is clear that Model 3 contains a much higher number of Type II errors than any of the other models. This is true regardless of how far into the future predictions are made. Despite the high Type II errors, Model 3 boasts some of the lowest Type I errors. This seems to suggest that Model 3 is not adequately learning the factors that contribute to bank failure. Model 4, by comparison, is has error values that are on par with the first two models.

5.2 PCA vs Risk Scoping Variables

When comparing Models 1 and 2, which used Principal Components as independent variables, to Models 3 and 4, which utilized rick scoping variables, we find that the models using PCA are much better at predicting failure. Models 1 and 2 were able to maintain their predictive capabilities up to two years out from our in-sample data. The risk scoping variables, on the other hand, had a higher rate of inaccurate predictions.

However, the higher accuracy of Models 1 and 2 come with a trade off. As discussed earlier, Principle Components are hard to interpret. Even after rotating the components and interpreting them in terms of call report variables, we cannot be sure what exactly is contributing to our models' predictions. The Risk Scoping variables, on the other hand, are self-explanatory. This is definitely of great importance when deploying these models to catch failures in practice.

5.3 In Crisis vs Out of Crisis

Overall, we found that the effectiveness of our model is contingent upon the time period that we use for evaluation. In particular, we found that during times of crisis, such as the 2008 financial crisis, our model is able to accurately predict failures up to two years out from the in-sample period. Outside of the financial crises, however, the predictive capabilities of our model diminish rapidly.

5.4 Exceptions Regressions

When we use the number of red flags to predict future failure, we found that the number of red flags was positively correlated with failure for two quarters. After this two quarter period, the predictive power of the total number of exceptions decreases tremendously. In fact, our regression

shows that there is no significant correlation between the number of red flags that a bank has and its risk of failure three or more quarters into the future.

Furthermore, the number of red flags that a bank has a similar decay in predictive power when it comes to predicting future red flags. Similar to predicting failure, the total number of exceptions in a given quarter is positively correlated with the number of exceptions one or two quarters into the future. After that, however, current exceptions cannot predict future exceptions.

6 Discussion

In this paper, we developed models that can predict banking failure accurately during crisis period, which is of great practical help. When a crisis occurs, examiners can apply our models to warn and support banks that may fail a few quarters in advance to avoid bankruptcy. We find that the models performed quite well both in and out of sample. This model can be utilized especially during crisis periods when examiners or regulators need to make a quick decision as to which banks to help first.

However, it is difficult to predict failure outside a crisis period. In normal time, the number of failures is almost nonexistent compared to a crisis period. For example, during normal times some quarters had single digit banking failure. These low amounts of failed bank observations lead to high Type I and Type II error and make it impossible for some models to converge. However, this is also an indication of the strength in our examiners and regulatory institutions, especially during non-crisis times. The red flags, or indicators that warn a bank of potential future failure, have value in predicting banking failure at most 2 quarters out. We investigated the relationship between red flags predicting future red flags and concluded a strong interaction between the red flags and future failure up to 2 quarters. We believe most banks who receive red flags from regulators self correct quickly. For the regulators, these red flags have eliminated potential failures. The effective job that banking supervisors did in the normal times does not require advanced models.

While our model accurately predicts which banks will fail during a financial crisis, it does not predict when or why a crisis will occur. That issue should be the topic of future research. We hope that this paper can give examiners and regulators new ideas to prevent future bank failure.

7 Appendix

Table 12: Estimates of Red Flags Predicting Future Red Flags

Variables	Coefficient	Standard Error
Red Flags (t-1)	0.038	0.000
Red Flags (t-2)	0.004	0.000
Red Flags (t-3)	0.000	0.000
Red Flags (t-4)	0.006	0.000
Red Flags (t-5)	-0.003	0.000
Red Flags (t-6)	-0.001	0.000
Red Flags (t-7)	-0.001	0.000
Red Flags (t-8)	0.004	0.000
Red Flags (t-9)	-0.003	0.000
Red Flags (t-10)	-0.001	0.000
Red Flags (t-11)	-0.001	0.000
Red Flags (t-12)	0.001	0.000
Observation	304,747	
Wald chi2(12)	224,930.50	
Prob > chi2	0.0000	

Table 13: Estimates of Red Flags Predicting Future Failure

Variables	Coefficient	Standard Error
Red Flags (t)	0.040407866	0.019142049
Red Flags (t-1)	0.148412005	0.030643958
Red Flags (t-2)	0.108383376	0.034365478
Red Flags (t-3)	-0.0639659	0.035668313
Red Flags (t-4)	0.057999686	0.035477339
Red Flags (t-5)	-0.04446633	0.033985556
Red Flags (t-6)	-0.01117234	0.033788816
Red Flags (t-7)	-0.02160306	0.034178035
Red Flags (t-8)	-0.03084536	0.034225332
Red Flags (t-9)	0.046900777	0.032794697
Red Flags (t-10)	-0.08661849	0.033719376
Red Flags (t-11)	0.098472346	0.033322575
Red Flags (t-12)	-0.14128147	0.02320919
Observation	304,822	
Prob > chi2	0.00	
Pusdeo R2	0.1543	

Table 14: Discrete Time Hazard Model

Variables	Coefficient	Standard Error
Red Flags	0.101	0.005
Observation	317,367	
Prob > chi2	0.0000	
LR chi2(45)	851.27	
Pseudo R2	0.1169	

Table 15: T-test Results for PCA Variables

Variable Label	Variable Name	Mean of 0	Mean of 1	Standard Error	T-test value	P-value
Scores for component 1	fullfailpca1	0.000	0.537	0.396	-1.358	0.174
Scores for component 2	fullfailpca2	-0.001	1.967	0.146	-13.526	0.000
Scores for component 3	fullfailpca3	0.000	0.095	0.127	-0.751	0.453
Scores for component 4	fullfailpca4	0.000	-0.186	0.092	2.025	0.043
Scores for component 5	fullfailpca5	0.000	-0.657	0.080	8.164	0.000
Scores for component 6	fullfailpca6	0.002	-3.851	0.078	49.247	0.000
Scores for component 7	fullfailpca7	0.000	-0.175	0.075	2.327	0.020
Scores for component 8	fullfailpca8	-0.001	1.142	0.069	-16.588	0.000
Scores for component 9	fullfailpca9	0.000	0.437	0.066	-6.625	0.000
Scores for component 10	fullfailpca10	-0.001	0.895	0.059	-15.117	0.000
Scores for component 11	fullfailpca11	0.001	-0.932	0.057	16.289	0.000
Scores for component 12	fullfailpca12	0.000	0.095	0.056	-1.689	0.091
Scores for component 13	fullfailpca13	0.000	-0.507	0.056	9.073	0.000
Scores for component 14	fullfailpca14	0.000	-0.656	0.055	11.978	0.000
Scores for component 15	fullfailpca15	0.000	0.052	0.054	-0.962	0.336
Scores for component 16	fullfailpca16	0.000	-0.297	0.053	5.621	0.000
Scores for component 17	fullfailpca17	0.001	-1.060	0.047	22.578	0.000
Scores for component 18	fullfailpca18	-0.002	3.000	0.046	-65.175	0.000
Scores for component 19	fullfailpca19	0.001	-0.900	0.046	19.661	0.000
Scores for component 20	fullfailpca20	-0.001	1.602	0.045	-35.814	0.000
Scores for component 21	fullfailpca21	0.001	-1.123	0.044	25.353	0.000
Scores for component 22	fullfailpca22	-0.002	3.262	0.043	-76.630	0.000
Scores for component 23	fullfailpca23	0.000	-0.790	0.042	18.779	0.000
Scores for component 24	fullfailpca24	0.000	-0.640	0.042	15.415	0.000
Number of Observation for 0	982996					
Number of Observation for 1	612					

Table 16: T-test Results for Risk Scoping Variables

Variable Label	Variable Name	Mean of 0	Mean of 1	Standard Error	T-test value	P-value
Total Securities/ Assets (%)	InvScope1	20.114	8.914	0.716	15.645	0.000
NPAs/ Assets (%)	AQScope21	0.920	0.394	0.098	5.374	0.000
Total Past Due Loans/ Loans (%)	AQScope13	0.356	0.333	0.057	0.404	0.686
Loan Loss Reserves/ Total Loans and Leases (%)	AQScope4	1.451	4.294	0.059	-47.939	0.000
Current Year Non Homogenous loan $$	AQScope22	492,973.603	77,848.474	393,336.366	1.055	0.291
B3: Capital Conservation Buffer (%)	CapScope5	4.209	0.043	9.879	0.422	0.673
Gain on Securities/ Avg Assets (%)	EarnScope5	-0.001	-0.144	0.011	13.253	0.000
Net Interest Margin (%)	EarnScope2	2.276	1.483	0.256	3.094	0.002
Noninterest Expense/ Avg Assets (%)	EarnScope4	2.049	2.409	0.193	-1.860	0.063
Provision Expense/ Avg Assets (%)	EarnScope6	0.121	2.195	0.024	-86.111	0.000
Return on Average Assets (%)	EarnScope1	0.430	-3.349	0.116	32.524	0.000
216947 Borrowings / Assets (Calc.)	LiqScope15	0.254	0.140	0.099	1.154	0.248
K434 Core Deposits / Total Deposits (UBPR)	LiqScope16	82.390	70.772	1.175	9.889	0.000
Total Securities/ Assets (%)	LiqScope6	20.114	8.914	0.716	15.645	0.000
Long-term Assets/ Assets (%)	SensScope1	23.056	11.749	0.785	14.402	0.000
Short-Term Assets/Short Term Liabilities	SensScope13	519.350	72.967	1,832.432	0.244	0.808
Non-Maturity Deposits/ Assets (%)	SensScope2	476.514	110.616	2,407.650	0.152	0.879
Non-Maturing Deposits/Long-term Assets (%)	SensScope12	476.514	110.616	2,407.650	0.152	0.879
Number of Observation for 0	410308					
Number of Observation for 1	487					

References

Prezipped large download files list. URL https://www7.fdic.gov/sdi/download_large_list_outside.asp.

Applied data mining and statistical learning, 2021a. URL https://online.stat.psu.edu/stat508/lesson/6/6.3.

Applied multivariate statistical analysis, 2021b. URL https://online.stat.psu.edu/stat505/lesson/11.

David H Erkens, Mingyi Hung, and Pedro Matos. Corporate governance in the 2007–2008 financial crisis: Evidence from financial institutions worldwide. *Journal of corporate finance*, 18(2):389–411, 2012.

Michael McCracken and Serena Ng. Fred-qd: A quarterly database for macroeconomic research. Technical report, National Bureau of Economic Research, 2020.

Jonathon Shlens. A tutorial on principal component analysis derivation , discussion and singular value decomposition: Semantic scholar, Jan 1970. URL https://www.semanticscholar.org/paper/A-TUTORIAL-ON-PRINCIPAL-COMPONENT-ANALYSIS-%2C-and-Shlens/bde7bb9b7478a23133c4731e6948a2ee123a0991.

John Tatom and Reza Houston. Predicting failure in the commercial banking industry. *Networks Financial Institute Working Paper*, (2011-WP):27, 2011.