CS 365 Final Project: Bank Failures

Laya Dang
Boston University
pd03@bu.edu

Abstract

Banks declare bankruptcy when they are unable to meet demands of depositors and stakeholders. This study will look into what characteristics of banks could potentially indicate its failure and if classification models could be used to flag which banks are likely to fail. A logistic regression that uses days since establishment, total assets, total equity, total deposits, number of physical locations, return on assets, net income, and stock/non-stock type are all used to create an accurate predictor on bank failures. The second half of the paper examines if macroeconomic factors impact the likelihood of bank failures, and found that unemployment rate and inflation rates are significant in the number of bank failures a year.

1 Introduction

This project will explore bank failures in the United States. A bank failure is any closing of a bank by a federal or a state regulatory agency. This project applies to commercial institutions and savings agencies that are federally insured, so it excludes organizations in the informal sector.

The known main reasons that cause banks to fail are under-capitalization, liquidity, stability, and fraud. In other words, these banks were unable to meet its obligation to depositors and other stakeholders, resulting in their bankruptcy.

1.1 Motivation

In 2023, 5 bank failures resulted in USD 548,705 million loss in total assets. During the 2007-2008 financial crisis, 165 banks closed, all of which held a total of USD 544,497 million in total assets (unadjusted for inflation).

These failures cause major disruptions for both individuals and business owners. These interruptions contribute to economic recessions and/or downturns, which overall hinder economic growth for the country. Some long-term economic consequences include a higher unemployment rate (as witnessed in the Great Depression) and weakened consumer/investor confidence.

Studying indicators that lead to bank failures will help uncover new ways to approach risk management and prevent future crises. The main motivation behind this paper is to both prevent and be prepared for these bank failures by understanding what factors are the most significant in predicting potential crises.

1.2 Prior Work

In a 1999 study from Texas A&M University by Kolari et al. [4], researchers used a logistic regression and nonparametric trait recognition model to develop classification models. The paper found both models were accurate (rates from 95% to 100%), but the latter minimized Type I and Type II errors in more tests. There were 18 major independent variables, categorized as the bank's size, profitability, capitalization, credit risk, liquidity, liabilities, and diversification.

However, this paper was published before the 2008 crisis and only analyzed large banks (banks holding more than USD 250 million). The paper also only focused on the 1-2 years before the failure occured and did not take into consideration macroeconomic variables.

In a more recent study from 2020 by Momparler et al. [5], researchers used a fuzzy-set qualitative comparative analysis (fsQCA), which allows them to identify a combination of factors, rather than isolating the effects of individual variables. This approach is more common for social sciences, as specific case-orientated variables are often more important.

Momparler et al.'s paper used 30 variables detailing each bank's financial ratios (earning assets to total assets, yield on earning assets, etc.) from 156 banks over a 15 years (2001 to 2015). One significant set found was low earning assets to total assets and low loan loss allowance to total loans and leases, meaning when riskier assets make up more of a bank's share and risk coverage is low, then the change of bank failures are higher.

2 Datasets

2.1 Federal Deposit Insurance Corporation (FDIC)

Using the same quarterly updated data set as the one used in Momparler et al. [5], this project will also pull data from the Federal Deposit Insurance Corporation (FDIC), which is publicly available through an API. The set of failed banks only goes back to the year 2000 and consist of 568 banks, while there are a total of over 25,000 institutions with the data going back to 1938.

More specifically, in total, there are 27,825 banks in the database, with 4,614 active banks as of the time this paper is written, March 2024. Although the database keeps track of 142 data points per bank, for the purpose of keeping the model simple and applying what we hypothesize are signficant variables, only the ones below would be of focus for now:

Table 1: FDIC Variables

Name	Description
CERT	Unique number assigned to each institution
ACTIVE	Binary classifier indicating if bank is still active, or our dependent variable
BKCLASS	Classification of bank type (see GitHub for more details)
ESTYMD	Established date
ASSET	Total assets owned by bank
EQ	Total equity capital
DEP	Total deposits
OFFICES	Number of branch locations
ROA	Return on assets
NETINC	Net income
MUTUAL	Stock or non-stock ownership type

2.2 Federal Reserve Economic Data (FRED)

The Federal Reserve Bank of St. Louis provides public data on key macroeconomic variables, with most dating back to the 1960s. The major ones that this project will focus on is described in the following table:

Table 2: FRED Datasets

Name	Description
Real GDP per capita	average economic output per person
Percent change in CPI	inflation indicator using the change in general levels of prices for consumer goods and services
Unemployment rate	percent of the labor force that is jobless and actively seeking employment

3 Research Questions and Methodology

This project will start with a basic logistic regression, using the standard machine learning library in Python ($sklearn$) in attempts to answer the first question (1) what is the most significant feature in predicting bank failure?

Then, to narrow down the time frame and focus on only two cases (failures in 2023 and failures in 2008-2009), we will use qualitative comparative analysis through the $scpQCA$ Python library to answer the second question (2) what combinations of regulatory, economic, and market conditions lead to bank failures?

4 Theoretical Foundation

4.1 Logistic Regression Overview

The logistic regression model is a simple model for binary classification and builds a strong foundation for the project, as it involves fitting the data into an S-shaped function:

$$P(Y = 1|X = x_n) = \frac{1}{1+exp(-(\beta_0+\beta_1 x_1 ...+\beta_n x_n))} n$$

where y is a binary variable (1 for failed bank, 0 for non-failed bank), $x_1...x_n$ are independent variables (return on assets, total deposits, net income, etc.), β_0 is the y-intercept, and $\beta_1...\beta_n$ is our estimated relationship. The output $P(Y = 1|X = x_n)$ is interpreted as the probability the classification is 1, given the independent variables. [3]

The function can be "linearized" by the logit function into an equivalent expression:

$$ln(\frac{P(Y=1|X=x_n)}{1-P(Y=1|X=x_n)}) = \beta_0 + \beta_1 x_1 ... + \beta_n x_n$$

The goal of this model is to estimate coefficients β using maximum likelihood estimation (MLE). MLE, or equivalenting minimizing the log-likelihood aims to find the set of parameters that makes the observed data most probable, which can be shown in the following objective function:

$$\min_\beta \sum_{i=1}^{n} \left(-y_i \log(p_i) - (1 - y_i) \log(1 - p_i)\right) + r(\beta)$$

4.2 Assumptions

The assumptions for this models are that the data shows none or little multi-collinearity, observations are independent, and sample size is sufficiently large. [7]

The biggest assumption I am making is that there is little multi-collinearity, which states that input parameters in the dataset are independent of each other. If they show dependency, it is more difficult to isolate individual effects. This is particularly important when it comes to time series data, such as macroeconomic trends, where the assumption of independence often gets violated due to the presence of auto-correlation. Techniques such as Variance Inflation Factor (VIF) analysis [6] can be used to detect multi-collinearity, preventing underestimation of standard errors and overconfidence in the model's predictions.

For this model, I also assume that observations are independent, meaning the failure of one bank does not lead to failure of another, which in the real world, may not be true. I also assume sample sizes are large enough, with 568 failed banks and a total of 27825 recorded banking institutions.

If these assumptions are violated, it can often be addressed through transformation of variables, adding interaction terms, adjusting regularization, or using a different model that do not rely on the same assumptions.

4.3 Regularization

In the *sklearn* Python package, the commands are straightforward and all essential tools are included in the *LogisticRegression* package. [1] There are four choices to control the regularization term

through the penalty argument: none, L1, L2, or a combination of L1 and L2 (elasticnet). Adding these arguments will help reduce over-fitting the training data. [3]

L1 regularization (or LASSO regularization), adds a penalty equal to the absolute value of the magnitude of coefficients. This can lead to some coefficients being exactly zero, which means L1 regularization can also perform feature selection and discard redundant outcomes. This approach is typically done in cases where only a few variables are suspected to actually influence the outcome.

L2 regularization, or Ridge regularization, adds a penalty equal to the square of the magnitude of coefficients. This tends to shrink the coefficients evenly but does not necessarily bring them exactly to zero. This approach is typical when many variables have some influence on the outcome.

4.4 Performance Analysis

To analyze the performance of our classification model, we will use the receiver operating characteristic (ROC) curve [2], a common and simple method to understand our results. This curve plots true positive rates (in our class, model correctly predicts bank failure when it is a bank failure) against false positive rates (model predicts a bank failure, even when one does not occur).

This plot is then assigned a numerical value, specifically the area under the ROC curve (AUC), ranging from 0 to 1. This value provides us with an aggregate measure of performance across all classification thresholds, which is whether a predicted probability is classified as a positive or a negative outcome. Adjusting the threshold affects the balance between true positive rate, where lowering the threshold means that the model classifies more items as positive, which might increase both the number of true positives and false positives.

A direct way to analyze our model's performance is through a confusion matrix, which includes four key metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It is important to consider the context of our problem, where false negatives (Type II errors) are much less desirable. A false positive (Type I error) may lead to unnecessary panic over a potential bank failure, but a false negative is dangerous in unpreparedness and repercussions for bank users. In other words, it is better to be more cautious and risk-averse when it comes to bank failures.

5 Trials

The GitHub repository for this project can be publicly accessed here.

5.1 Variables

For the first few trials, we will use variables from Table 1 in Subsection 2.1. Later on, Federal Reserve variables (Subsection 2.2) will be considered.

5.2 Trial 1 Results

Using the *LogisticRegression* class from *sklearn*, I fit the training model (using an 8:2 training to testing data set ratio), and then ran predictions on the test set. Before that, to handle NaN entries (4045 rows had at least one), I dropped rows with more than 2 NaN values and filled the rest with 0.

Since the logistic regression returns the probability for $y = 1$, we can adjust the threshold, which by default, the threshold in this library accepts for $P \geq 0.5$. Further, without adjusting default arguments, the regression uses a mix between L1 and L2 penalty regularization term. With this set-up, we proceed to build a model that returns an 81.7% accuracy on the test data, with the following confusion matrix:

Table 3: Confusion Matrix for Trial 1

		ACTUAL	
		Positive	**Negative**
PREDICTED	**Positive**	3841	26
	Negative	846	52

Although this deceptively appears to be a good model based purely on the confusion matrix and accuracy, the Type I error (False Positives) is 33.3% while Type II error (False Negatives) is 18.0%, which is comparatively high. We could improve this by changing the threshold, which can be seen in the ROC curve plotted below, using a built in function from *sklearn*.

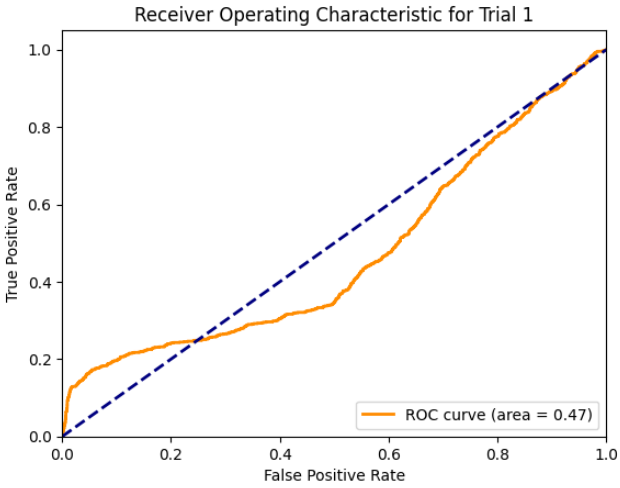

Figure 1: ROC Curve for Trial 1

We are provided that our AUC value is 0.47, which is considered very poor, as 0.50 is the expected value for a random choice. This result is an indicator that our model is even performing somewhat inversely, or worse than a random guess. Ideally, we would want a high true positive rate and a low false positive rate for every threshold, but based on the results, it seems like the false positive rate also increases as true positive rate increases.

In this first run, I wanted to run a logistic model on all the variables from my hypothesis to then weave out variables that may be less significant. This run includes 11 variables with this trial, but includes the 10 dummy variables that Bank Institution Class *BKCLASS* classification variable creates, making a total of 23 variables. Initially, I thought bank classification is significant, as commercial banks, saving banks, federal banks, state banks, and so on, all operate under different models and regulations. However, the 10 classifications under this variable may be too granular and case-specific to each point in our data set.

In Trial 2, we expect that removing Bank Institution Class *BKClass* and decreasing the number of variables will simplify the model and improve our performance.

5.3 Trial 2 Results

After dropping *BKCLASS*, and having 10 variables left, we run another Logistic Regression with the same train/test sets as before. The accuracy improved to 95.1%, and we get the confusion matrix as shown below in Table 3.

Table 4: Confusion Matrix for Trial 2

		ACTUAL	
		Positive	**Negative**
PREDICTED	**Positive**	3788	79
	Negative	154	744

The true positive rate is similar to before (around 18%), but the true negative rates improved drastically, with the Type I error now being 9.60% and Type II error being 3.9%, much more ideal compared to results in Trial 1. We get the corresponding ROC curve.

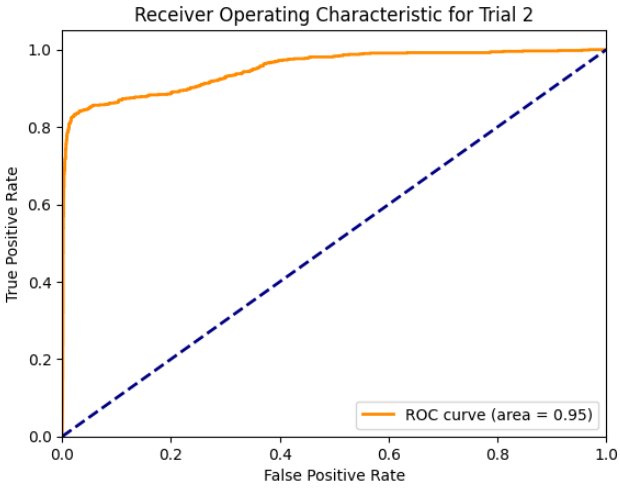

Figure 2: ROC Curve for Trial 2

The AUC value is now 0.95, which is a great improvement compared to our previous model. This indicates our model has a high level of ability to distinguish between the two classes, and there is a 95% chance that the model will be able to correctly distinguish between a randomly chosen positive instance and a randomly chosen negative instance.

To examine what is happening in our model, we can look at each coefficient β.

Table 5: Coefficients for Trial 2

Feature X_n	Value
ASSET	$-4.719084e-08$
NETINC	$-9.744106e-07$
EQ	$-7.396067e-07$
DEP	$1.462343e-07$
ROA	$6.573426e-01$
MUTUAL	$-1.065300e-02$
OFFICES	$1.832630e+00$
AGE (difference between current date and ESTYMD)	$-1.036482e-04$

The relationships are all similar to what I would expect, where a dollar unit increase in assets, net income, equity, etc. would decrease the odds of failure. These small numbers are also expected, as we are examining at single dollar increases relative to banks that hold assets in million dollars.

What seems to be surprising is that the coefficient for number of offices *OFFICES* is positive at around 1.83, suggesting that more number of physical locations is correlated with the probability of failure. In particular, when a bank builds one additional location where clients are able to open accounts, there is a greater risk of failure suggested. This could be a larger area of research and a strong incentive for banks to invest in online banking services rather than having clients come in-person.

Since *sklearn* does not come with statistical tests for the coefficients (standard errors, p-value, confidence intervals), we cannot determine which coefficients are the most significant to the model.

6 Additional Trials

For the next part of this research that asks what combinations of regulatory, economic, and market conditions lead to bank failures, I feel a machine learning model is not necessary, but rather tests for statistical significance is more crucial to our understanding.

6.1 Variables

Isolating only failed banks from our previous dataset, we focus on closing dates. Then, we use real GDP per capita, percent change in CPI (inflation indicator), and unemployment rate as indicators of economic performance around the time of bank failure. We combine factors by year, testing the relationship between number of bank failures per year from 2000 to 2023, and the average of all the macroeconomic variables previously mentioned.

6.2 Results

As shown below, this is the correlation matrix between bank failures and macroeconomic indicators of unemployment rate, real GDP per capital, and inflation rate, taken as an average per year.

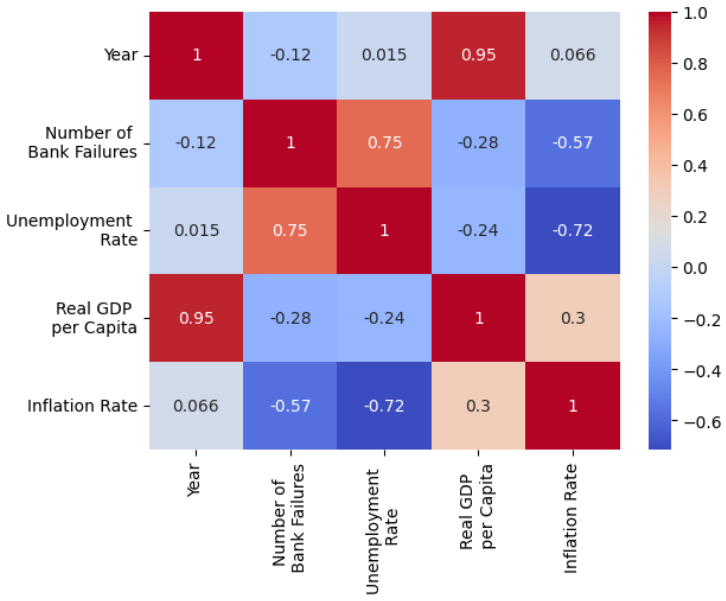

Figure 3: Correlation Matrix for Bank Failures vs. Macroeconomic Indicators

This alone does not provide with much detail, so shown below are correlation coefficients along with it's p-value from the pearson correlation coefficient test (using *pearsonr(col A, col B)* from Python's *scipy* library).

Table 6: Coefficients and p-value for Number of Bank Failures vs. Macroeconomic Indicators

Macroeconomic Indicator	Correlation Coefficient	p-value
Unemployment Rate	0.75	0.00052
Real GDP per Capita	-0.28	0.28
Inflation Rate	-0.57	0.017

From basic statistics, we know that lower p-value means greater statistical significance between the observed difference (our current result vs. null hypothesis), with values less than 0.05 considered statistically significant. This tells us unemployment rate and inflation rate are significant when considering its impact on the number of bank failures per year, while real GDP per capita is not. Positive unemployment rates tend to increase number of bank failures, while lower inflation rate increases.

This matches with what I had expected. Real GDP per capita alone is known to not be a good sole indicator of the status of the economy, as it is output and production focused, rather than measuring quality of life and spending confidence in a country's citizens.

One thing to consider with this model, however, is that number of bank failures could further impact indicators. For example, if unemployment rates are already high and multiple bank fails, unemployment rate would further increase, and cause a spiral.

7 Conclusion

7.1 Challenges and Issues

Although I did have more granular data for the macroeconomic analysis to go by month and year, rather than just year, I ran into issues when aggregating the data. This ended up with many months where no banks fail, and too much emphasis on one bank failure in a month. Therefore, I found that combining data by average per year is better for this approach, but have not tested the alternative.

7.2 Potential Future Research Areas

For future trials, if I decide to continue with using the logistic regression model and improve, I first perform t-tests on each feature coefficient and determine which ones are significant. I would also test more on different penalty options (LASSO, Ridge, or none) to understand how that could affect my model. Using different training and testing divisions may also help me understand how well my model would really perform.

Macroeconomic indicators could also potentially be incorporated within the logistic regression model as well, with some sort of indicator that recognizes that a bank who has previously overcome a period of macroeconomic hardship is more likely to overcome it again.

References

[1] "1.1.11. Logistic Regression." scikit-learn: machine learning in Python, 2024. https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

[2] "Classification: ROC Curve and AUC." Google for Developers, July 18, 2022. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[3] Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong. (2020). *Mathematics for Machine Learning*. Accessed March 2024. https://mml-book.com.

[4] Kolari, J., Glennon, D., Shin, H., & Caputo, M. (2002). Predicting large US commercial bank failures. *Journal of Economics and Business, 54*(4), 361–387. https://doi.org/10.1016/s0148-6195(02)00089-9

[5] Momparler, A., Carmona, P., & Climent, F. (2020). Revisiting bank failure in the United States: a fuzzy-set analysis. *Economic Research-Ekonomska Istraživanja*, 33(1), 3017–3033. https://doi.org/10.1080/1331677x.2019.1689838

[6] Pardoe, Iain. "10.7 - Detecting Multicollinearity Using Variance Inflation Factors." STAT 462 Applied Regression Analysis: Detecting Multicollinearity Using Variance Inflation Factors, 2018. https://online.stat.psu.edu/stat462/node/180/.

[7] Shai, Ben-David, and Shalev-Shwartz Shai. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Accessed March 2024. https://www.cs.huji.ac.il/