# Assessment of the financial health of US banks using AI/ML

## \*\*Author\*\*

Pravin Powale

## Executive summary

The goal of this project is to leverage machine learning techniques to assess the financial health of US banks using quarterly financial ratio data publicly available from the FDIC. The analysis utilizes datasets containing financial ratios for US banks from 2001 to 2023, information about insured banks as of June 30, 2014, and a list of failed banks. By clustering banks based on their financial ratios, the project aims to categorize them into four groups: Extremely Healthy (Blue), Healthy with Low Risk (Green), Medium Risk (Orange) and High Risk (Red). This indirect approach involves using known healthy banks and failed banks to label the clusters accurately.

Principal Component Analysis (PCA) has been employed to reduce the number of input features to key features, followed by the application of Logistic Regression. This model was evaluated using metrics like False Positives, False Negatives, Confusion Matrix and accuracy score. As part of the next steps, additional classification techniques such as Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Randon Forests may be employed to improve the classifier performance. This project is significant as it explores the potential of AI/ML techniques in bank regulation, offering a more efficient alternative to traditional rule-based systems. If successful, these techniques could be extended to other regulatory domains, enhancing the monitoring and understanding of the overall health of the economy.

## Rationale

Why should anyone care about this question?

One of the key processes of bank regulation is to monitor the financial health of banks under supervision. This has a direct impact on the understanding of the overall health of the economy with wide range of global implications. "Business Rules" based systems have been used by the financial regulators in the past. This is an attempt to explore the potential of applying AI/ML techniques to solve this problem, as the rule based systems have to re-configured, re-programmed frequently to ensure acceptable levels of performance. If successful, the techniques can be potentially extended to other regulatory domains.

## Research Question

What are you trying to answer?

Every quarter FDIC publishes 700+ financial ratios of every US bank and makes it available publicly. The main hypothesis is that the financial health of a bank is reflected in these financial ratios, which can be used to build AI/ML models to classify banks into various categories of financial health.

Since labeled data (with financial health categories) is not available publicly, an alternate approach of labeling the quarterly bank data is implemented. This consists of first performing a Principal Component Analysis (PCA) and reducing the dimensionality of the data, by conducting Elbow Analysis of the Sigma array generated during Singular Value Decomposition (SVD). A K-Means clustering with n\_clusters = 4 is then performed, to form the clusters of banks. The cluster numbers are then used to label each bank. This labeled data is then used to train a Logistic Regression Model which can be used to predict labels (in other words Financial Health Categories) of banks based upon financial ratios for another quarter.

This submission has considered four categories of financial health. Since the cluster numbers do not directly represent the financial health of banks, a standard candle (in other words a well known bank) will be used to determine the "Blue" cluster (high financial health) and the "Red" cluster (High financial Risk). The remaining two clusters will need to be inspected manually or compared with another source of data to identify which one is Green (Low Risk) and Orange (Medium Risk).

## Data Sources

What data will you use to answer you question?

1. US Banks Financial Quarterly Ratios 2001-2023 – located at <https://www.kaggle.com/datasets/neutrino404/all-us-banks-financial-quarterly-ratios-2001-2023>

This dataset has quarterly data of financial ratios for all the US banks that have existed since 2001. The data was downloaded from FDIC website and processed further to computed ratios using the percentages provided for columns ROA, ROAQ, ROAPTX, ROAPTXQ, ROE, and ROEQ.

The dataset contains 700+ ratios and the following key attributes

* REPYEAR = 'REPORT YEAR'
* REPDTE = 'The last day of the financial reporting period selected.'
* STNAME = 'STATE NAME'
* NAME = 'INSTITUTION NAME'
* CERT = 'FDIC Certificate Id. A unique NUMBER assigned by the FDIC used to identify institutions and for the issuance of insurance certificates'

1. U.S. FDIC Insured Banks and Financial Institutions Asset Data. This was downloaded using FDIC's BankFind Suite: Find Institution Financial & Regulatory Data

The download link (that inludes the date range is)

<https://banks.data.fdic.gov/bankfind-suite/financialreporting/report?establishedEndRange=2%2F18%2F2025&establishedStartRange=01%2F01%2F1792&inactiveEndRange=2%2F18%2F2025&inactiveStartRange=01%2F01%2F1970&incomeBasis=YTD&institutionType=banks&limitEstablishedDate=false&limitInactiveDate=false&pageNumber=1&reportPeriod=20230630&resultLimit=25&sortField=CERT&sortOrder=ASC&unitType=%24>

This dataset contains information about the insured banks in the United States. It includes information about each bank, including their name, address, assets size, and classification code. The data is as of 2023-06-30. The data record has the CERT Id which is the same identifier as in the data source 1. See NOTES below.

1. Additionally, the dataset “Failed Banks List” available at https://www.fdic.gov/bank-failures/failed-bank-list will be used to label a cluster of banks created using the financial ratios. See NOTES below.

### NOTES

1. The dataset 2 has not been used directly in the notebook, but the downloaded excel file was used to filter banks on Asset Size >= USD 10B, which was then uploaded as CERT\_List\_20230630.csv
2. The dataset 3 has not been used to locate any failed banks in any of the clusters, but none were found. This is probably due to the fact none of the banks selected (Asset size >= USD 10B) have failed so far. As a result, all the 4000+ banks will be considered for analysis in the next steps.

## Methodology

What methods are you using to answer the question?

The following steps were executed:

1. Clean the data source 1 to drop columns that have NaNs. This reduced the columns from 796 to 770
2. Select bank data corresponding to a quarter
3. Drop numeric columns "CERT", "REPDTE" and "REPYR" which can cause noise during data analysis. The number of columns now equal 767.
4. Drop non-numeric columns. Only one was found in the cleaned data, which NAME representing bank name. Number of columns now = 766.
5. Normalize the financial ratios by first subtracting the mean and then dividing the result by standard deviation
6. Drop the columns that contain NaNs after normalization. This can be a result of standard deviation being 0 causing division by 0. The number of columns now equal 699.
7. Perform PCA on the 766 dimensional data after initial cleaning. This resulted in 158 Principal components.
8. Perform Elbow Analysis and keep only the components to the left of the Elbow. This reduces the number of Principle Components used to 29.
9. Perform K-means clustering on the 29 dimensional data, with n\_clusters = 4. Use the cluster numbers to label the banks in each cluster.
10. Use the dataframe before application of PCA as X. Use the labels generated as y. Perform train\_test\_split on the original dataset to create
11. Create a LogisticRegression classifier (lgr) and fit the data (X\_train, y\_train)
12. Use lgr to calculate the accuracy, confusion matrix, false positives and false negatives.

## Results

What did your research find?

The following bank clusters were created

1. cluster 3 consists of 10 banks (Seems to be Blue) as it has the following banks.

**bankname** **CERT**

STATE STREET BANK&TRUST CO 14

PACIFIC WESTERN BANK 24045

GOLDMAN SACHS BANK USA 33124

**BANK OF AMERICA NA 3510**

**WELLS FARGO BANK NA 3511**

HSBC BANK USA NATIONAL ASSN 57890

**JPMORGAN CHASE BANK NA 628**

BANK OF NEW YORK MELLON 639

CITIBANK NATIONAL ASSN 7213

NORTHERN TRUST CO 913

1. cluster 0 consists of 8 banks (Seems to be Green) as it has the following banks

**bankname** **CERT**

FIRST-CITIZENS BANK&TRUST CO 11063

SYNCHRONY BANK 27314

**AMERICAN EXPRESS NB 27471**

**DISCOVER BANK 5649**

BARCLAYS BANK DELAWARE 57203

COMENITY CAPITAL BANK 57570

ALLY BANK 57803

SALLIE MAE BANK 58177

1. cluster 1 consists of 55 banks (Red or Orange) - needs to be analysed further
2. cluster 2 consists of 86 banks banks (Red or Orange) - needs to be analysed further

## Classifier results

The LogisticRegression classifier yielded an accuracy of 0.925 on the test data. The notebook contains the Confusion Matrix and false positives and false negatives calculations.

## Next steps

What suggestions do you have for next steps?

1. Perform analysis of all 4000+ banks available in the dataset. This will potentially include the failed banks which will help determine the Red cluster.
2. Use additional classification algorithms such as KNeighborsClassifier, DecisionTreeClassifier, SVM and potentially Random Forests to determine and achieve higher accuracy scores.
3. Identify each cluster's health (Blue, Green, Orange, Red) by either manual analysis of key banks in the cluster or some independent source of information which needs to be explored. At present, the Blue and Green clusters seem to be well defined according to the general knowledge about the health of the banks in these clusters and due to relatively low counts of banks (10 and 8 respectively) in these clusters.

## Outline of project

### Link to notebook 1

<https://github.com/pravinpowale/berkeley-AI-ML-Capstone-Assignment-20.1/blob/main/BDA%20-%20Notebook%202.ipynb>

### Link to GitHub Repository

<https://github.com/pravinpowale/berkeley-AI-ML-Capstone-Assignment-20.1>

## Contact and Further Information

Pravin Powale

email - pravinpowale@yahoo.com