Capstone Project

Banks’ Financial Health Assessment using AI/ML Final Report

1. **Define the Problem Statement:** (*Provide a well-defined problem statement that clearly articulates the goals, challenges, and potential benefits of your machine learning solution.)*

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, it poses a challenge to use supervised learning models such as Logistic Regression for bank health classification.

To circumvent this problem, 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 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). A failed bank data file downloaded from FDIC public site is used to identify the “Red” and “Orange” clusters, where “Orange” indicates Medium to High financial Risk. The 4th cluster by elimination is designated as “Green”. In the analysis done for this project, neither the “Green” nor the “Blue” cluster contained any banks from the failed banks data file.

Note that, once a bank fails, FDIC stops reporting its financial ratios for future quarters. As a result, the quarterly data for quarter ending 20230630 contained four banks which eventually failed. The two banks that failed within one year after 20230630 were in cluster #2 which was marked as “Red”. The other two banks failed more than one year after 20230630, were in cluster #1 which was marked as “Orange”. This has been explained within the code in detail. **It must be emphasized here that there were several hundred other banks that were in clusters #1 and #2. These banks according to the data analysis were at elevated risk of failure unless risk mitigation strategies were employed. Obviously, this requires further research outside of the scope of this Capstone project.**

1. **Model Outcomes or Predictions:** (*Identify the type of learning (classification or regression) and specify the expected output of your selected model. Determine whether supervised or unsupervised learning algorithms will be used.)*

After creating the labeled data, the following classification models were created and evaluated:

1. Logistic Regression
2. Ridge Classifier followed by GridSearchCV analysis
3. Support Vector Machines (SVM) classifier followed by GridSearchCV analysis
4. Random Forest classifier followed by GridSearchCV analysis

For each GridSearchCV analysis performed, the best model for that classifier was determined. It was found that the RidgeClassifier with the following parameters was the best model.

**Best RidgeClassifier(alpha=10000.0, class\_weight='balanced', random\_state=42)**

**Best Ridge Model Accuracy: 0.9779474130619169**

**Best Ridge Model Recall: 0.9887315010570824**

**Note that All classifiers used class\_weight="balanced" to account for imbalanced cluster data.**

A revised Ridge Model was built with the same parameters but with limited number of features or ratios. This was necessary as the Bank Data Columns can vary from quarter to quarter including the quality of data in the columns. There may be some new columns in a quarter as well as some columns may be absent, when compared to the quarter for which the classification model was built. A SHAP analysis of the best Ridge Model was done and 90 percentile SHAP values were used to select the features for the revised model. This approach decreased the likelihood of data related issues in the model. This has been clearly explained in the Python notebook.

1. **Data Acquisition:** (*The deliverable at this step is to identify what data you plan to acquire and use with your model. For the best results, data should come from multiple sources and your analysis for including specific data should be clear. Please provide a clear visualization to assess the data’s potential to solve the problem as well.)*
2. 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. Additionally, the dataset “Failed Banks List” available at https://www.fdic.gov/bank-failures/failed-bank-list has been used to label a cluster of banks as “Red” or “Orange” created using the financial ratios. This has been explained with the Python notebook.
2. **Data Preprocessing/Preparation:** (*For this deliverable, you are tasked with detailing how you cleaned the data for your notebook.)*
3. What techniques did you use to ensure your data was free of missing values, and inconsistencies?

* Clean the data source 1 to drop columns that have NaNs. This reduced the columns from 796 to 770
* Select bank data corresponding to a quarter
* Drop numeric columns "CERT", "REPDTE" and "REPYR" which can cause noise during data analysis.
* Drop non-numeric columns. Only one was found in the cleaned data, viz. NAME representing bank name.
* Normalize the financial ratios by first subtracting the mean and then dividing the result by standard deviation.
* Drop the columns that contain NaNs after normalization. This can be a result of standard deviation being 0 causing division by 0.
* **The quarter ending 20230630 was selected for the PCA and Clustering. The revised Ridge Model built on this data was used to perform predictions on the quarter ending 20221231. This could have been the other way, but was done to reuse already developed code and save some valuable time.**

1. How did you split the data into training and test sets?

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y)

Used all default values. On second thought, random\_state = int could have been used to give consistent splitting of data across various runs.

1. Please include any necessary analysis and encoding steps you took as well.
2. **Modeling:** (*For this deliverable, please document your selection of machine learning algorithms that you selected for your problem statement from the first deliverable.)*

The modeling has been described in detail in sections 1 (Define the Problem Statement) and 2 (Model Outcomes or Predictions) above. The Python code is also well documented to describe the various models used.

1. **Model Evaluation:** (*Share your model evaluation here. What types of models did you consider for your problem (classification, regression, unsupervised)?  Articulate the evaluation metrics you used and how you determined which model was most optimal for your problem.)*

As described above, PCA and K-Means clustering was used with elbow analysis performed on the PCA results. This is described in the notebook. The PCA reduced the number of features from 704 original features to 58 new features (linear combinations of the original features).

The cluster or labeled data created was used to develop different classification models as explained above. For each classification model, its accuracy and recall were computed. Since failure of a bank can be potentially catastrophic, a model with high Recall value was selected to minimize false negatives.

Finally, the revised Ridge model was applied to quarter ending 20221231 data and predictions created. It is interesting to note the following:

* Cluster #0 banks such as
  + JPMORGAN CHASE BANK NA / CERT = 628
  + WELLS FARGO BANK NA / CERT = 3511

Continued in to be in Cluster #0

However, BANK OF AMERICA NA / CERT = 3510 which was in cluster #0 now appeared in cluster #2 which is the “Red” cluster. This can either be a false positive or a case to be explored deeper.

* Similarly, all the failed banks appeared in clusters #1 (Orange) or #2 (Red) depending upon when they failed, except for the following two banks:
  + SILICON VALLEY BANK / CERT = 24735 failed on 3/10/2023
  + SIGNATURE BANK / CERT = 57053 failed on 3/12/2023

These two banks failed within less than a quarter after 20221231 and were identified as cluster #0 (Blue). According to news articles both these banks failed suddenly and unexpectedly. This highlights a weakness of the approach considered here. A potential was to address this issue could be looking at some derived features such as Rate Of Change (ROC) across the adjacent quarters to build the model. The complexity of this approach is outside the scope of the Capstone project.

1. **Conclusion and Next Steps:**

The set of models considered in the presented analysis are based on the hypothesis that the 700+ financial ratios’ data contains information regarding the financial health of individual banks. The clustering analysis performed (in the absence of publicly available labeled data, along with identification of “**standard candles**”[[1]](#footnote-1) for each cluster seems to be working at first sight but still needs additional validation of results from the experts in the financial regulatory sector. In addition to considering additional features such as quarterly Rate Of Change (ROC) of certain ratios, one can split the data across large, medium and small banks which may require different sets of financial ratios to be considered for each type of bank. Similarly, a much more complex time series analysis of specific banks of interest can be performed to predict a small number of features that directly determine the financial health of a bank.

Another possible future variation can involve use of clustering algorithms such as DBSCAN, which generate the appropriate number of clusters on their own. However, this can pose a challenge as the number of clusters may vary from quarter to quarter, which will require further analysis.

Finally, this hypothesis can be generalized to extend to quarterly financial data reported by corporations to determine their financial health.

1. The term is borrowed from astronomy, which refers to an astronomical object with a known luminosity, like Cepheid variables, or like Type Ia supernovae that serve this purpose. These standard candles help measure distances to star clusters and galaxies by comparing their known luminosity to their observed brightness. Here the use of the term is much more limited than in astronomy. [↑](#footnote-ref-1)