GEORGETOWN UNIVERSITY
The Graduate School of Arts & Sciences

Predictive Modeling of Small Bank Survival Under Increasing Federal Funds Rate

Yunhan Zhang*[1], Yihan (Nathen) Bian†[2], and Qiaojuan (Tina) Tu‡[3]

[1] *Department of Data Science and Analytics„ Graduate School of Arts and Sciences„ Georgetown University„ Washington, DC 20057*
[2] *Department of Data Science and Analytics„ Graduate School of Arts and Sciences„ Georgetown University„ Washington, DC 20057*
[3] *Department of Data Science and Analytics„ Graduate School of Arts and Sciences„ Georgetown University„ Washington, DC 20057*

Dated: April 23, 2023

Faculty Advisor:
Prof. Dr. Nakul R. Padalkar

Final Report Submitted for:
CSBS 2023 Annual Data Analytics Competition

Table of contents

*Electronic address: yz1048@georgetown.edu
†Electronic address: yb214@georgetown.edu
‡Electronic address: qt34@georgetown.edu

Abstract

Due to the global impact of COVID-19, the United States Federal Reserve System (the Fed) has repeatedly increased the Federal Funds Rate, the interest rate for overnight borrowing between banks, to help cool down the high inflation since March 2022. As the country strives toward recovery, examining how the banking industry behaves during this rapid increase in interest rates is crucial. The primary objective of this research paper is to use machine-learning techniques to analyze how banks managed their risk exposure during the historical interest rate hikes from 2004 to 2008 and the survival probability of banks with varying financial structures.

Introduction

The 2023 Conference of State Bank Supervisors (CSBS) Annual Data Analytics Competition allows the participating team to investigate how banks are impacted during a rapidly rising interest rate environment. To better understand the behavior of the banking industry during a rapidly growing interest rate environment, it is crucial to examine historical interest rate hikes and how different banks navigated them. We can see in Figure 1 that the rising interest rates starting in 2022 are quite similar to what happened to start in 2004. For this reason, we focus on investigating the behavior of community banks in 2004-2008, more specifically, what are some of the key aspects that make a community bank survive or not during a high-interest rate period. Therefore, our hypothesis posits that the statistical characteristics of banks will exhibit a comparable alteration due to the increment in interest rates. To test this hypothesis, we intend to implement regression models, classification models, and neural network models utilizing the FFIEC Call Reports Data obtained from the Federal Financial Institutions Examination Council (FFIEC) as inputs [2]. Through this model, we aim to predict the survival duration with a community bank's given report data scenario. The final model will be utilized to forecast the expected lifespan of community banks that submit reports in 2022.

Data Sources

In accordance with the rules of this competition, participating teams are suggested to use bank Call Report data from FFIEC (Federal Financial Institutions Examination Council) and FDIC (Federal Deposit Insurance Corporation), and other publicly available data sources. Accordingly, the primary data utilized in this research paper are:

A. FFIEC Call Reports Data Reporting Year 2004-2007 and 2022

Call reports data is the primary dataset that this research paper use, and it contains financial information reported quarterly by banks and financial institutions in the United States. The data covers over 600 various financial and accounting metrics, including balance sheet items,

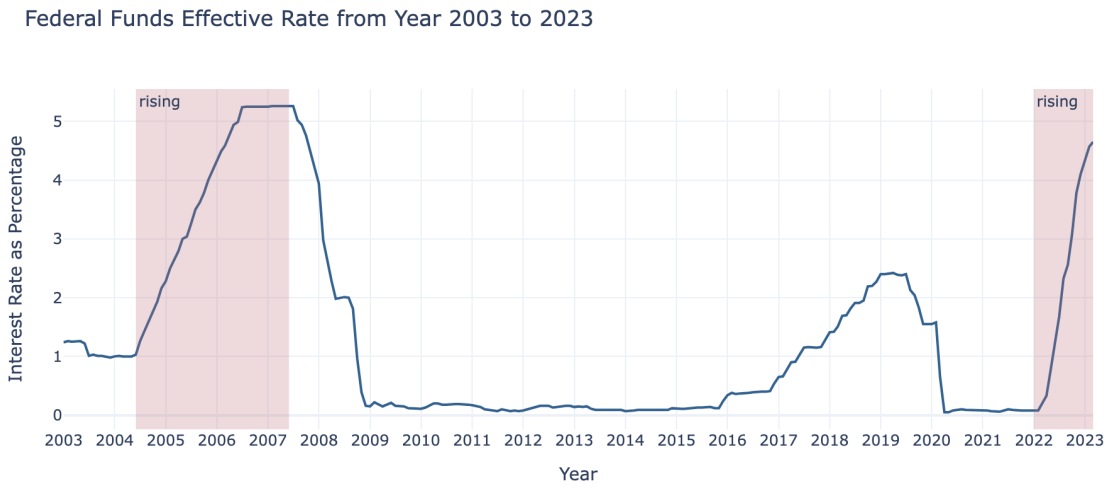

Figure 1: Federal Funds Effective Rate Two Rapid Rising Periods

income statement data, loan portfolios, and other relevant financial ratios. Each quarter, approximately 8000 banks in the US, including both community and national banks, file this report with either form FFIEC 031 or FFIEC 041 [2].

B. FDIC Community Bank Reference Data Year 2003-2009 and 2017-2022

This data is used with the Call Reports Data to filter out institutions marked as community banks. This data contains a variable called "cb" that indicates whether a bank belongs to the community bank category. More specifically, non-community banks reported in Call Reports Data will be filtered out to match the goal of this data analysis [1].

Project Workflow Preview

The project workflow in Figure 2 shows the general workflow of the project.

Data Cleaning

The primary data source for this project is the FFIEC CDR Call Report data for each bank. We used the historical community banking reference data to filter the relevant institutions. Given the similar historical situation that occurred between 2004 and 2008, we decided to combine the Call Report data from those years to serve as our training and validation dataset.

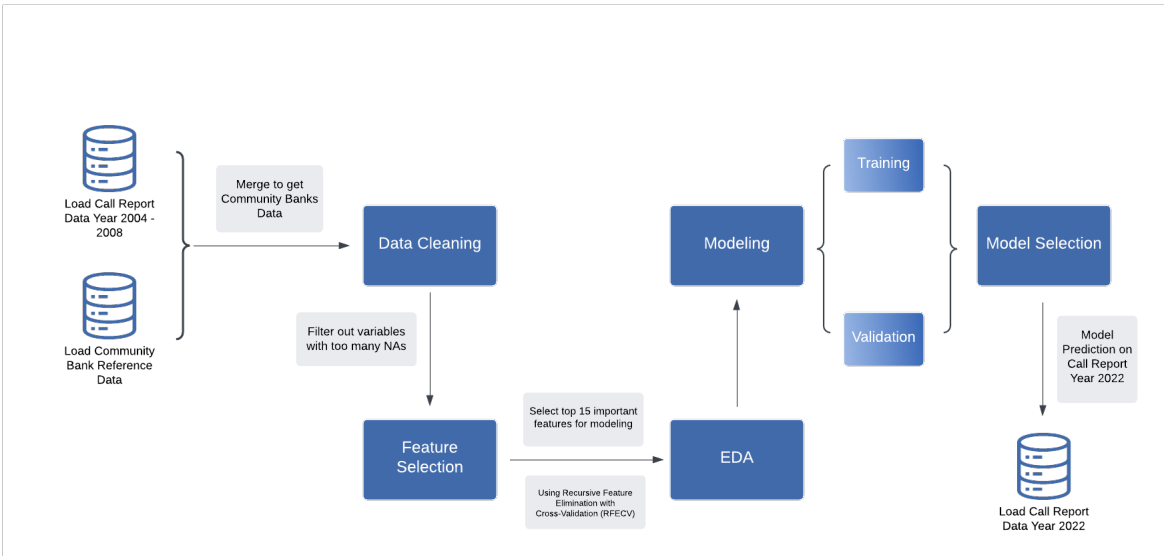

Figure 2: Project Workflow

Labeling

In order to facilitate our further data analysis, we created a new variable named "date_index" which assigns a numerical label to each of the 16 quarters in the dataset. Specifically, the first quarter of 2004 is labeled as "1", the second quarter as "2", and so on.

Handling Missing Values

During the initial data inspection, we observed significant numbers of columns with a high proportion of missing values. Since our ultimate goal is to create a reliable predictive model, columns with a large number of null values cannot be used in the analysis. To maintain the data's integrity and reliability, we set a threshold of 90% for null values in each column. Any column exceeding this threshold was dropped from the dataset. After this process, the data is left with around 150 variables [3].

Filtering Community Banks

We used the historical community banking reference data to filter out the relevant community banks from the call report data. By performing a merge operation on the datasets using the unique bank identifiers as the key, we created a new dataset containing only the financial and regulatory information of community banks for the years 2004-2008. This dataset will be used for training and validation purposes in the development of our predictive model.

In conclusion, the data-cleaning process for this project involved handling missing values by dropping columns with over 90% null values and filtering community banks using the historical community banking reference data. The resulting dataset, which covers the years 2004-2008, is now ready for use in the development of a predictive model to determine the survival or failure of community banks during periods of Federal Reserve rate hikes.

Feature Selection

Next, we selected the most important features for building the traditional machine learning model. The feature selection process involves several steps to filter and refine the data before identifying the most relevant features for the model.

First, we create a list of feature columns, excluding the ones used to identify the community banks. We then filter the training dataset based on specific quarter index values to focus on the relevant periods. We remove columns with more than 10% missing values to ensure data quality. Afterward, the feature columns list was updated to exclude the dropped columns.

To handle the remaining missing values, we fill them with 0, which is a reasonable default value for our dataset. We then select a subset of the training data to be used for feature selection. This is done by sampling a fraction of the dataset, using a random state for reproducibility.

With the dataset prepared, we perform feature selection using Recursive Feature Elimination with Cross-Validation (RFECV). We employ a time series cross-validation strategy with five splits to maintain the temporal structure of the data. The `RandomForestRegressor` is used as the base estimator for RFECV, with a negative mean squared error scoring metric.

We extract the feature importances from the estimator upon fitting the RFECV model to the training data subset. The features are then sorted based on their importance, and the top 15 features are selected for use in the traditional machine learning model. Selecting the most relevant features, we can create a more accurate and efficient the predictive model to assess the viability of community banks during periods of Federal Reserve rate hikes.

Pre-Processing for Modeling

After selecting the top 15 features, we must prepare the data for the subsequent modeling process. The selected features are flattened, creating a row for each selected variable at quarters 1, 6, 11, and 16. This adjustment allowed the sliding window technique to capture temporal dependencies and patterns in sequential data during the modeling process.

Finally, imputation is performed on the merged dataset containing the flattened features and target variables. For each relevant column in the dataset, the median is computed, and missing values are replaced with the median if the corresponding last value is greater than or equal to the date_index of the column. The remaining missing values are filled with zeros. This

Table 1: Selected Feature Description From FFIEC 031 or FFIEC 041

Name	**Importance (%)**	**Description**
RCONB835	0.04669	Loans to depository institutions and acceptances of other banks
RIAD4180	0.04235	Expense of federal funds purchased and securities sold under agreements to repurchase
RCON3505	0.03867	Debt securities and other assets (exclude other real estate owned and other repossessed assets)
RCON2150	0.03491	Other real estate owned
RCON5400	0.03295	Loans secured by real estate: Revolving, open-end loans secured by 1 to 4 family residential properties and extended under lines of credit
RCON3499	0.03219	Loans secured by real estate: Secured by multifamily (5 or more) residential properties in domestic offices.
RCON2930	0.03161	Other liabilities
RCON6558	0.03158	Loans to finance commercial real estate, construction, and land development activities (not secured by real estate)
RCONB528	0.03010	Loans and leases held for investment
RCONB576	0.02873	Loans to individuals for household, family,and other personal expenditures:
RIAD4513	0.02272	Interest expense incurred to carry tax-exempt securities, loans, and leases acquired after August 7,1986 , that is not deductible for federal income tax purposes
RCON3495	0.02264	Loans secured by real estate: Secured by farmland in domestic offices
RCON2160	0.02083	Other assets
RCON5369	0.02058	Loans and leases held for sale
RIAD4301	0.01984	Income (loss) before applicable income taxes and discontinued operations

process ensures that imputations are performed based on the institution's survival status and data availability at specific time points.

Exploratory Data Analysis (EDA)

In this section, we present the results of an exploratory data analysis (EDA) conducted on the dataset of community banks. The primary objective of this EDA is to gain insights into the trends and patterns in the number of community banks over time. By understanding these trends, we can better comprehend the factors affecting the growth and stability of the community banking sector, which can be valuable information for future research and policy decisions.

Before analyzing the data, we prepared the dataset by sorting it according to the reporting period end date and mapping each unique date to an index. This enabled us to systematically compare the community bank data across different time periods.

The data analysis focused on two aspects: the persistence of the community banks from the initial time period (2004) in subsequent time periods, and the total number of community banks in each time period. These comparisons allowed us to investigate the stability of the community banking sector over time and understand its growth patterns.

To visualize the trends in the data, we created a plot displaying the number of community banks from the initial time period (2004) that persisted in each subsequent time period, along with the total number of community banks in each period. This visualization helped us observe the changes in the community banking sector over time, providing a comprehensive understanding of its growth and stability.

The EDA revealed valuable insights into the changes in the number of community banks over time. We examined the trends and found that the community banking sector experienced growth and stability variations throughout the investigated period. These insights can inform future research and policy decisions related to community banks and the broader financial industry.

Modeling and Analysis

The ultimate goal of the model is to provide community banks a tool to provide their survival possibility since they may not have done similar professional stress tests as the big banks regularly did. We want to find the best model that could fit our dataset. We will first go over our dataset. The data consists of fifteen core features with four time stamps and two target variables, "survived" and "last." The variable "survived" is a binary series, where a "0" indicates failed to survive, and "1" indicates survived. The variable "last" is a numerical value ranging from 1 to 20, indicating how long it lasts from the beginning time. The model will use the features to predict if one bank survived and for how long. The model will output "last" for

"20" for survival since it is the maximum value of ‘last‘ in the data set. It needs to be noted that there is a significant imbalance in the dataset since the majority of banks survived the 2007-2008 crisis. A complete modeling diagram is presented in Figure 3.

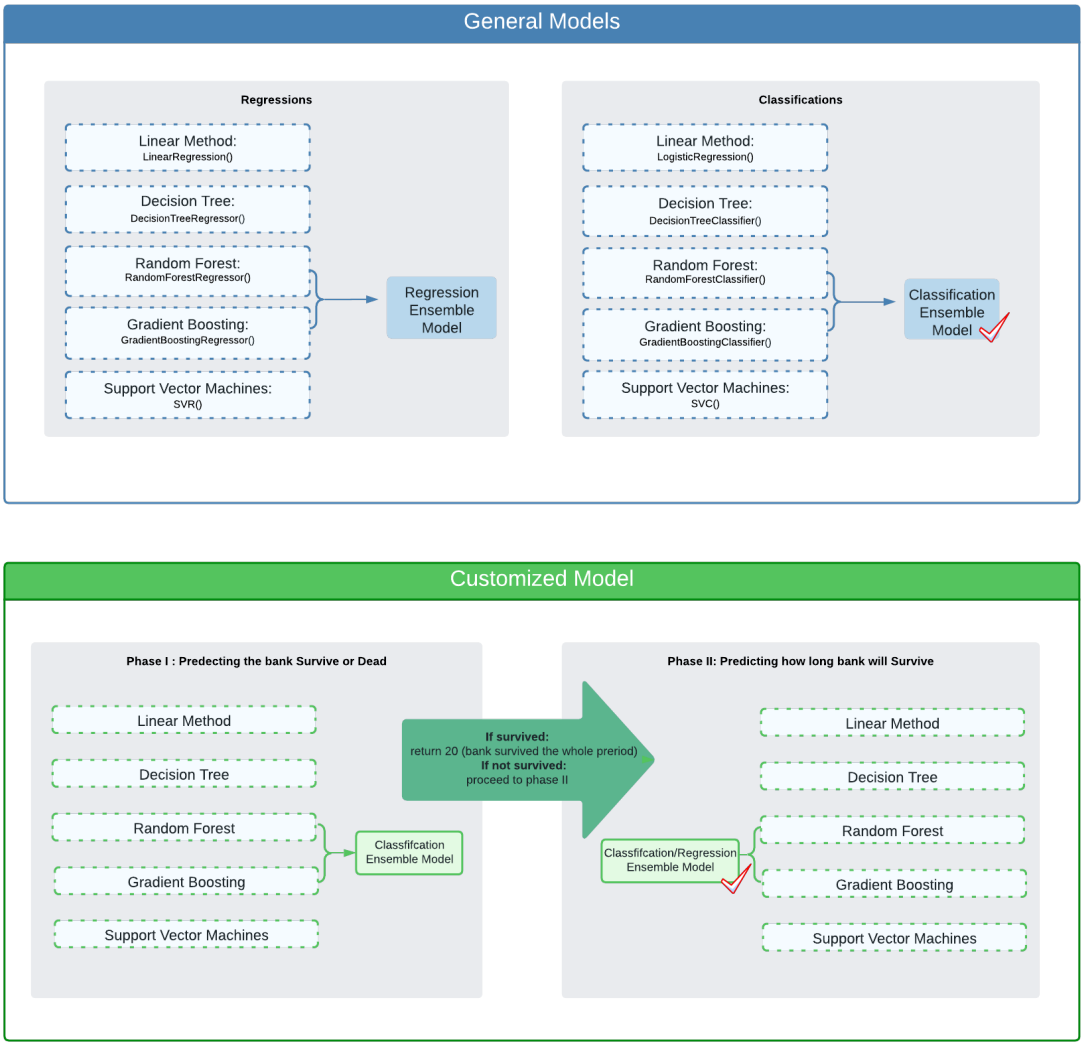

Figure 3: Modeling Workflow Diagram

We first want to use the existing general model to predict failure. Since our last variable can be both interpreted as a discrete number series or a unique 20 labels variable, we will utilize both regression and multi-classification models.

Regression Models

In this study, we performed a train/validation split on the given dataset. We explored a variety of base models for regression, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Regression (SVR). Each model was trained on the training set (X_{train}, Y_{train}) and evaluated using the mean squared error (MSE) calculated from their respective predictions on the validation set (X_{val}, Y_{val}) [3].

Subsequently, we selected two models, Random Forest, and Gradient Boosting, to create an ensemble model utilizing the VotingRegressor method. The ensemble model combines the strengths of multiple models to improve prediction accuracy potentially. It was trained on the same training data, and its performance was evaluated using the MSE metric. The MSE results for all models were reported in Table 2

Table 2: Mean Squared Error for different models

Model	**Mean Squared Error**
Linear Regression	19.5673
Decision Tree	0.9980
Random Forest	0.5666
Gradient Boosting	0.5211
SVR	19.0712
Ensemble Model	0.5328

Although the MSE values seemed promising, upon plotting the real values against the predicted values. The predicted values seem clustered at 3, 8, and 13. This suggests that even the best regression model does not fit this dataset satisfactorily. Further investigation into alternative models is required to improve model performance. Thus we want to look at the classification models.

Multi-Classifier Models

In this analysis, we employed several base models for classification, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Classification (SVC). Each model was trained on the training set (X_{train}, Y_{train}) and evaluated using the accuracy metric calculated from their respective predictions on the validation set (X_{val}, Y_{val}).

Furthermore, we selected two models, Random Forest, and Gradient Boosting, to create an ensemble model utilizing the VotingClassifier method from scikit-learn. This approach combines the strengths of multiple models to enhance prediction accuracy, specifically for classification models potentially. The ensemble model was trained on the training data, and its performance was evaluated using the accuracy metric. We chose the top three models with the highest accuracy for visualization. The accuracy results for all models are reported as follows:

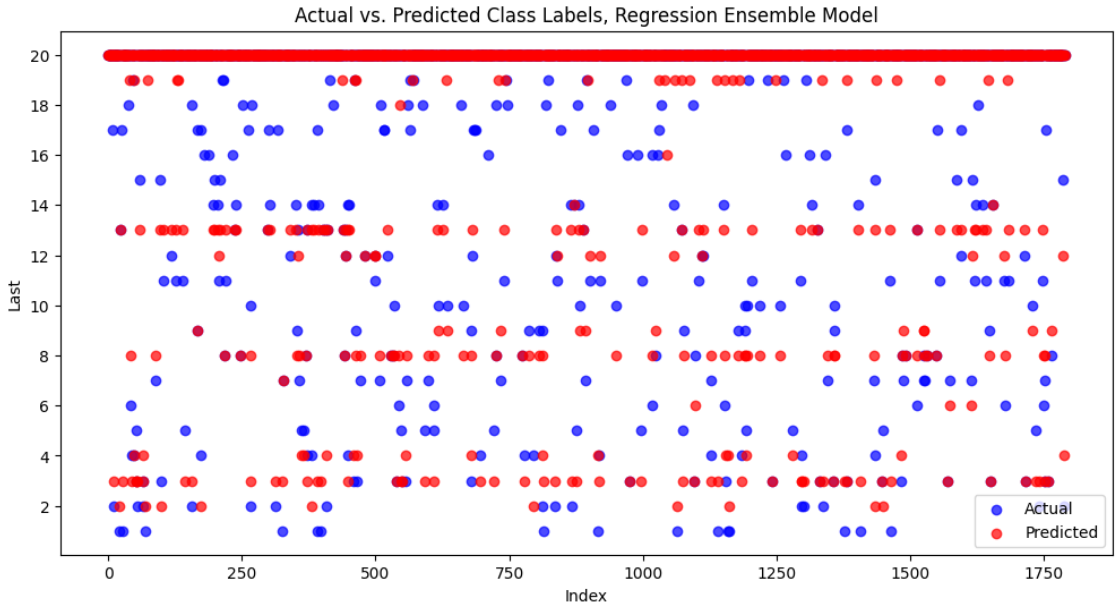

Figure 4: Ensemble Model Prediction on Validation Set

Table 3: Model Accuracy

Model	**Accuracy**
Logistic Regression	0.8653
Decision Tree	0.8367
Random Forest	0.8723
Gradient Boosting	0.8437
SVC	0.8402
Ensemble Model	0.8430

Random Forest Classifier

As seen from the graph and the diagonal of the confusion matrix, this classifier performs well for this dataset. However, it should be noted that the classifier does not predict any values above quarter 15, which is not an accurate representation.

Gradient Boosting

With the Gradient Boosting Model, we observe more random red points throughout the plot. This indicates that the predictor does not avoid predicting any values, making it a better choice

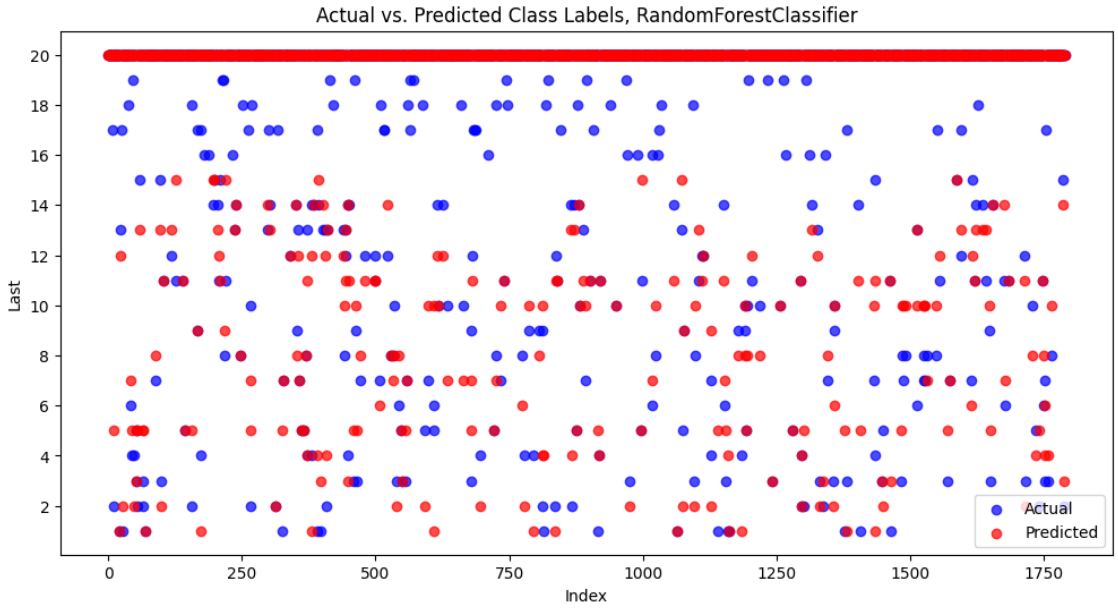

Figure 5: Random Forest Classifier Prediction on Validation Set

than the Random Forest classifier. However, it has a slightly lower accuracy compared to the Random Forest classifier.

Ensemble Model

We aimed to increase the overall accuracy using the Random Forest classifier while not sacrificing the data above 15 years by incorporating the Gradient Boosting classifier in an ensemble model. Unfortunately, the results were not very promising, with slightly lower accuracy and reduced prediction accuracy on survival compared to the Random Forest classifier.

General Models Summary

Upon visualizing the predictions, we observed that the data points were scattered across the plot, indicating higher accuracy than regression models. Analyzing the models' performance, it becomes evident that the Gradient Boosting classifier is the best model for this dataset, as it has slightly higher accuracy and demonstrates a more spread-out pattern on the plot, avoiding any prediction biases. Based on these findings, we conclude that the Gradient Boosting classifier is the most suitable choice for this dataset.

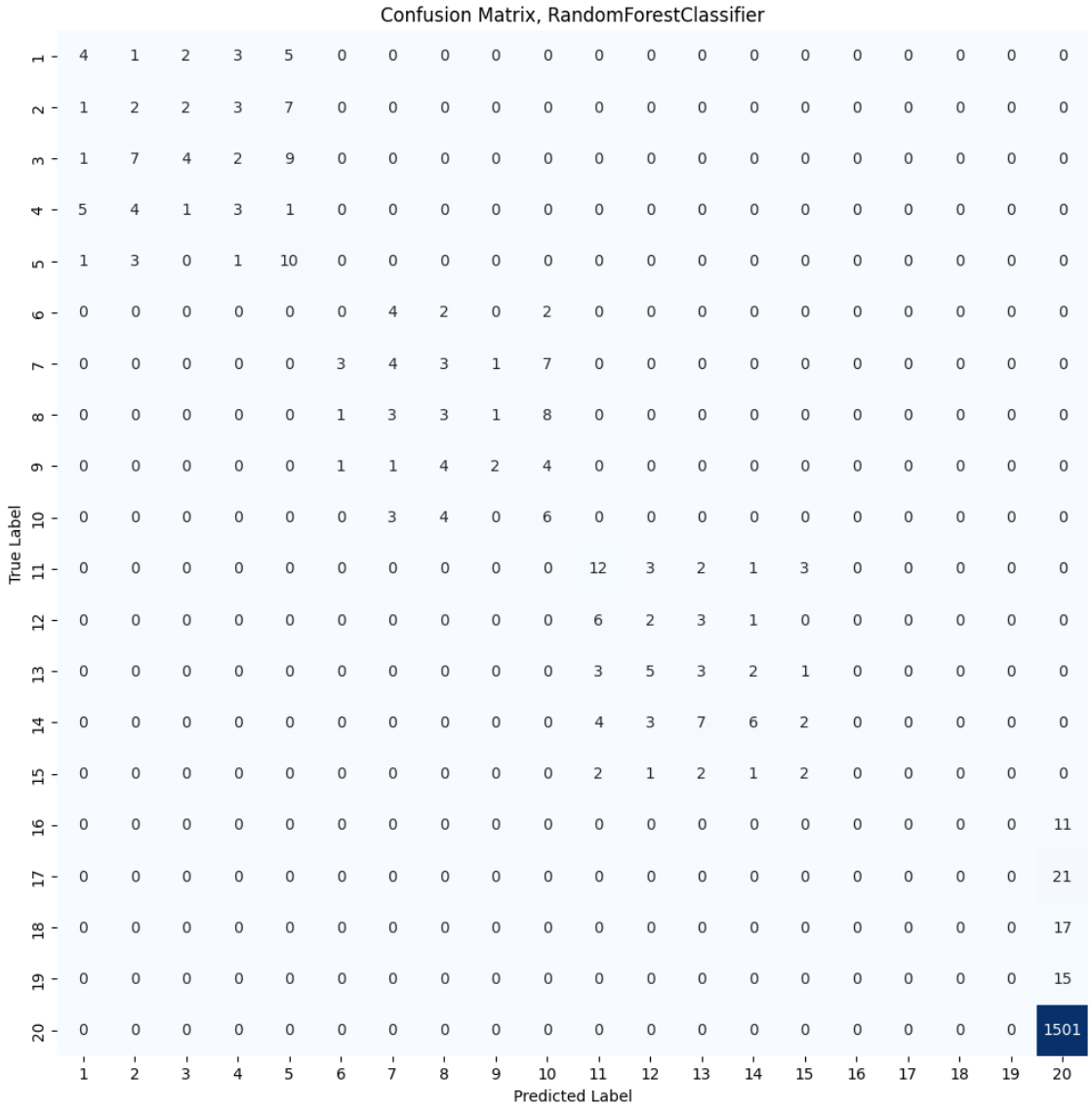

Figure 6: Random Forest Classifier Prediction Confusion Matrix on Validation Set

Customized Models

Motivated by the imbalanced dataset and previous work, we propose a two-phase modeling approach. The data processing is divided into two parts: The first phase employs a suitable model to predict if the bank survives; if the bank survives, the maximum last value is output. Otherwise, the data is fed to the second phase, which predicts the bank's duration, ranging from 1 to 19. This approach is expected to mitigate the effects of the imbalanced dataset and create a more robust model.

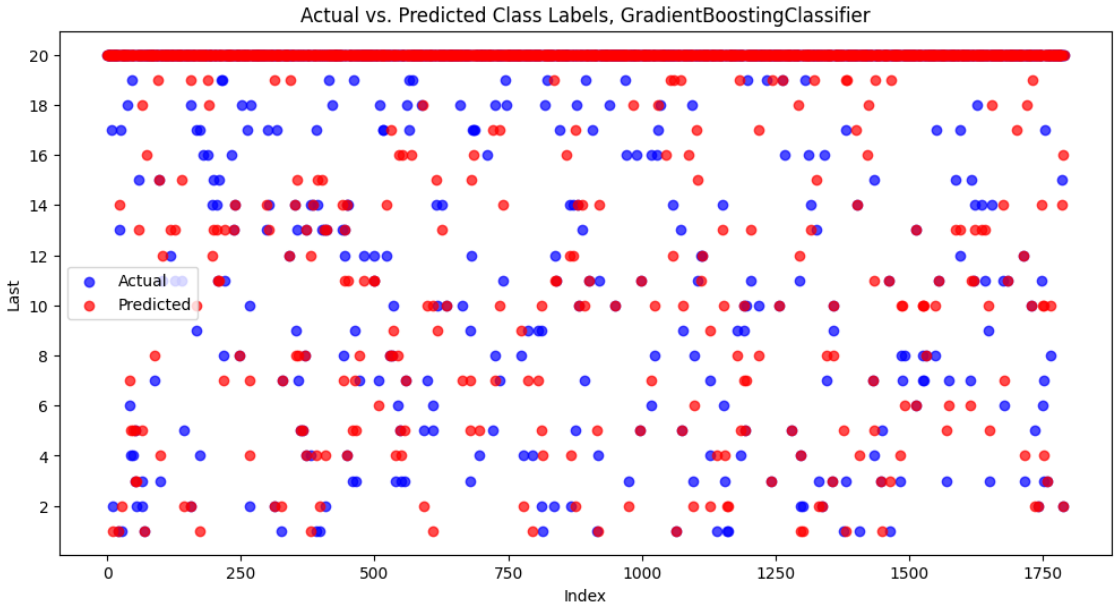

Figure 7: Gradient Boosting Prediction on Validation Set

Phase I: Binary Classifier

We seek the optimal classification model for predicting survivability. A procedure similar to previous searches is performed, yielding promising results displayed in the table below:

Table 4: Model Accuracy

Model	**Accuracy**
Logistic Regression	0.9546
Decision Tree	0.9267
Random Forest	0.9714
Gradient Boosting	0.9714
SVC	0.8723
Ensemble Model	0.9714

Random Forest, Gradient Boosting, and Ensemble Model (consisting of Random Forest and Gradient Boosting) all have the same accuracy. To enhance the model's robustness for future unseen data, we used the ensemble model to construct the customized model.

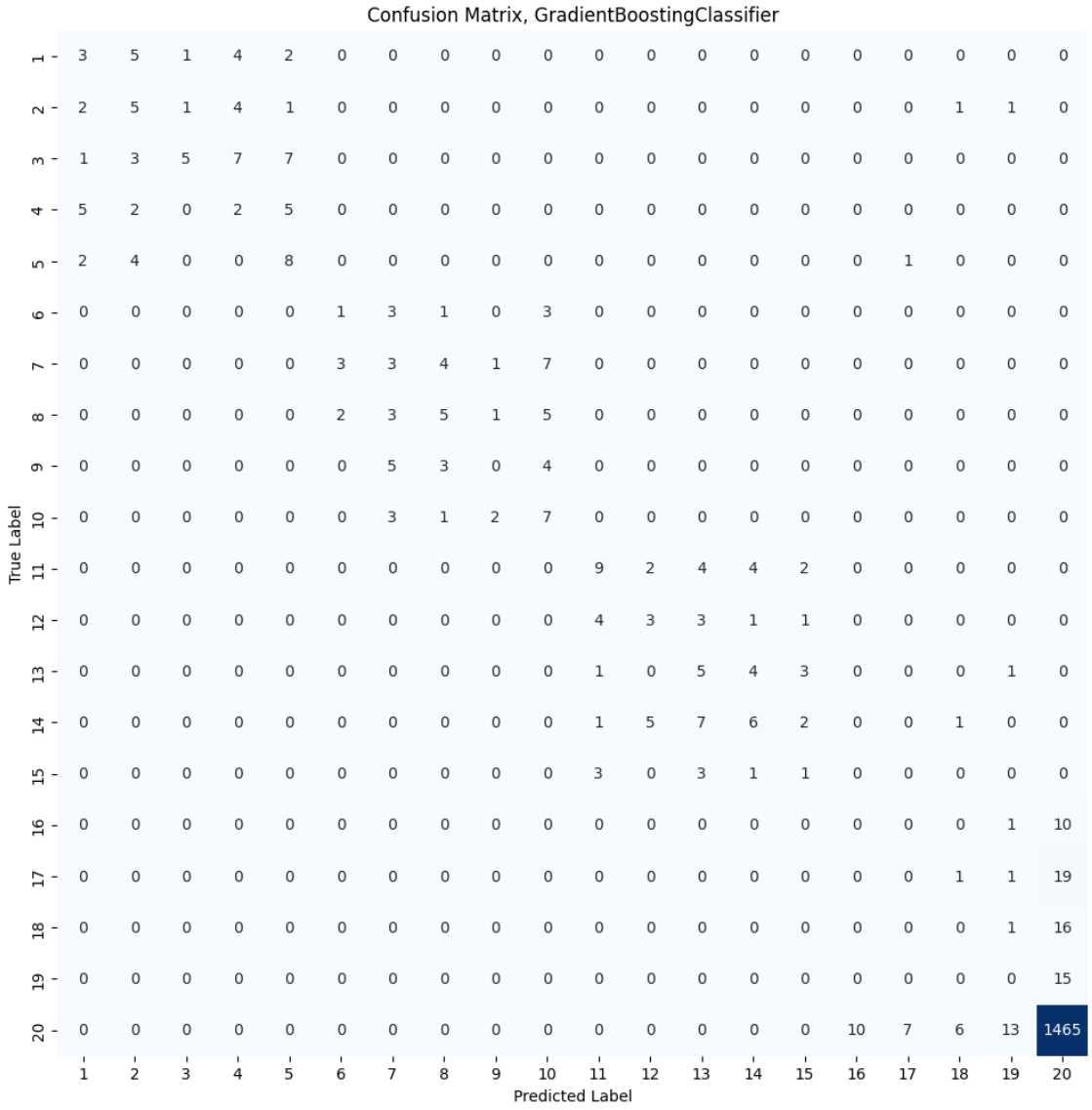

Figure 8: Gradient Boosting Prediction Confusion Matrix on Validation Set

Phase II: Predicting Survival Duration

To train this model, all data with $last == 20$ are removed to improve accuracy. We follow the same procedure as before to identify the best model for this predictor. The mean squared error (MSE) and accuracies for the regression and classification models are shown in the tables below:

We see similar clustering with the regression methods. Thus, this model set is forfeited, and

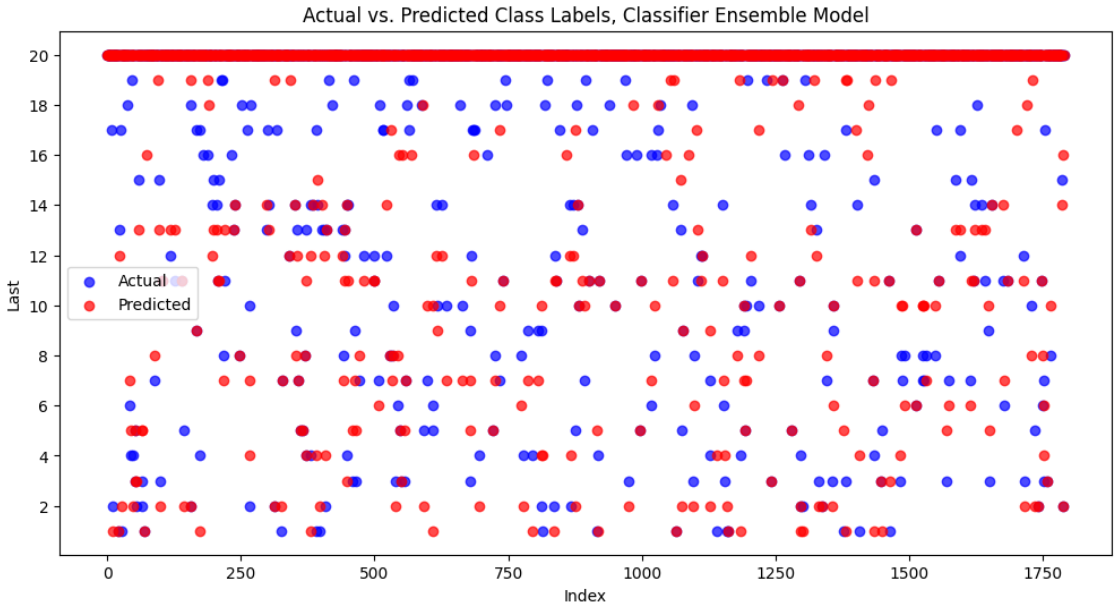

Figure 9: Ensemble Model Prediction on Validation Set

Table 5: Regression Models' Mean Squared Error

Regression Models	**Mean Squared Error**
Linear Regression	37.4609
Decision Tree	3.5427
Random Forest	2.0526
Gradient Boosting	1.9085
SVR	21.3178
Ensemble Model	1.9350

further rounding accuracies are not provided.

The classification model performs better than anticipated. The predicted points are dispersed across the graph, and the red dots, while not precisely on the blue actual value dots, are relatively close. This indicates that the actual error of this method is not as low as the accuracies presented above, as we are dealing with numerical predictions rather than categorical ones. Therefore, we choose the ensemble model of Random Forest and Gradient Boosting to construct the combined model.

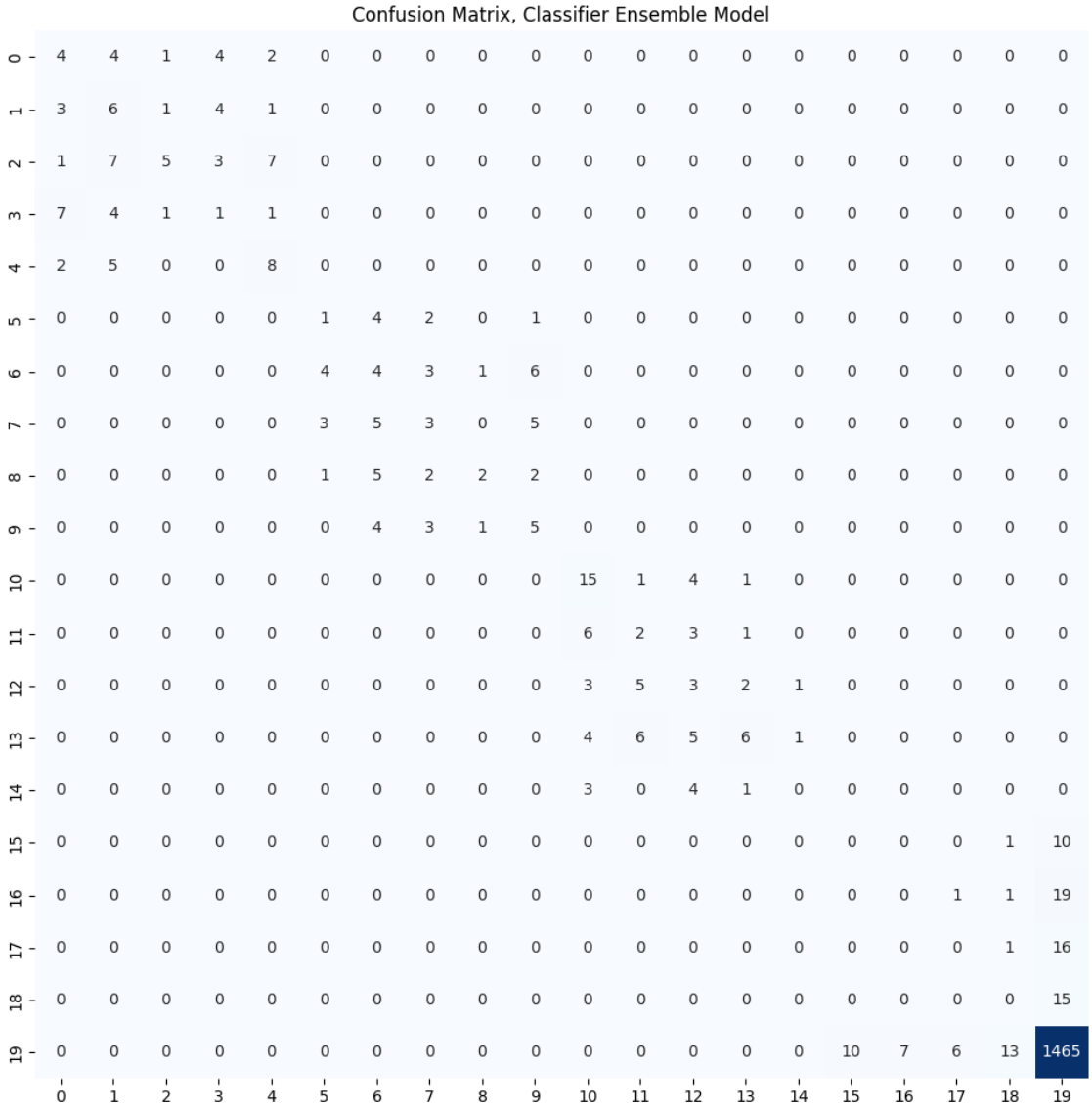

Figure 10: Ensemble Model Prediction Confusion Matrix on Validation Set

Final Model Performance

We proposed a CombinedModel that combines two classifiers, explicitly leveraging the RandomForestClassifier and GradientBoostingClassifier from scikit-learn. This approach aims to enhance prediction capabilities by training two models: one for survival prediction and the other for time duration. We utilized the VotingClassifier method to create ensemble models for both classifiers.

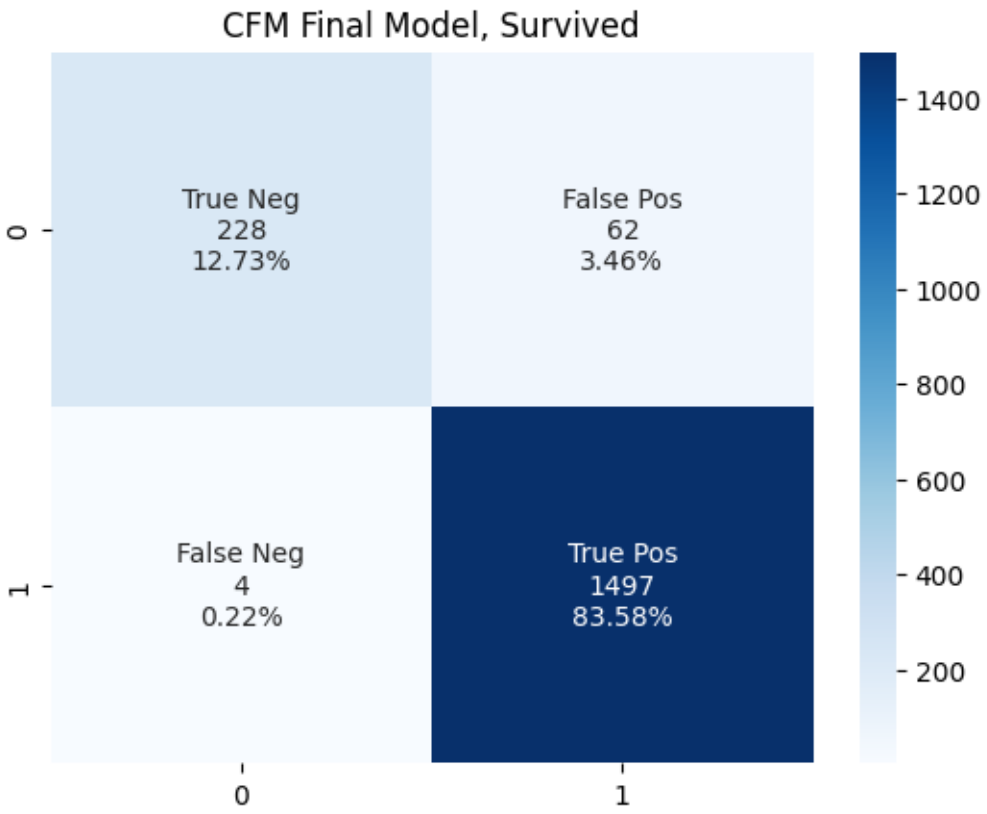

Figure 11: Ensemble Model Prediction Confusion Matrix on Validation Set

To implement the CombinedModel, we initialized two VotingClassifier models consisting of the selected RandomForestClassifier and GradientBoostingClassifier. We then created and fit the CombinedModel using the training data for survival and the training data for time duration. The CombinedModel predicts the survival and duration by combining the predictions of both classifiers, resulting in a final prediction.

These performance results indicate that the CombinedModel provides a more robust and accurate solution by combining the strengths of RandomForestClassifier and GradientBoosting-Classifier, thus making it a suitable choice for the given dataset. By addressing the imbalanced dataset through a two-phase modeling approach and employing ensemble models, our customized model demonstrates improved prediction capabilities, contributing to its potential effectiveness in real-world applications.

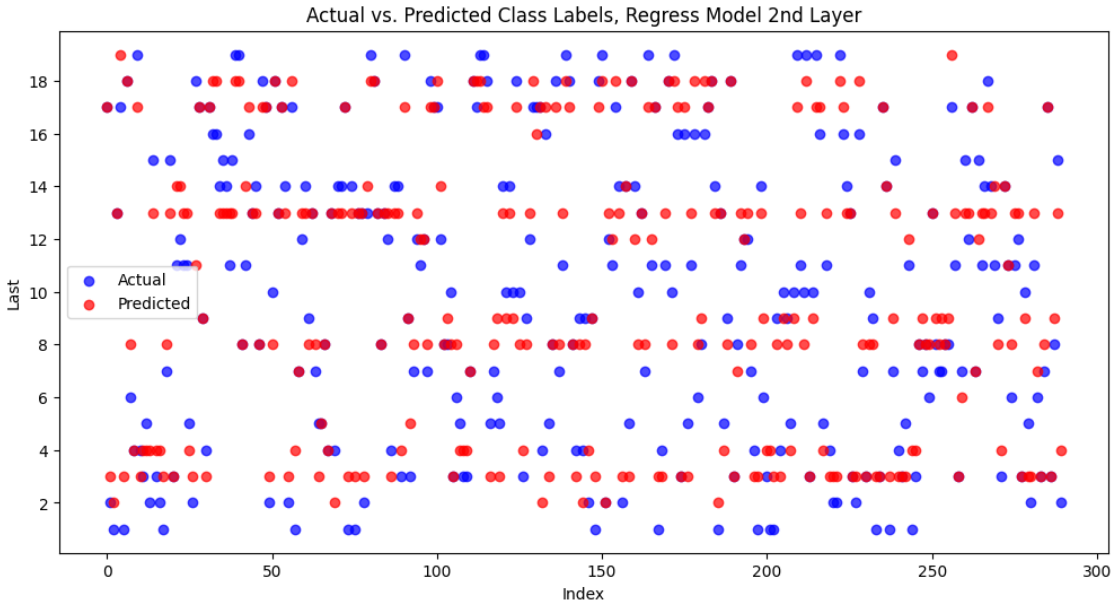

Figure 12: Ensemble Regression Model Survival Duration Prediction on Validation Set

Table 6: Classification Models' Accuracy

Classification Model	Accuracy
Logistic Regression	0.2265
Decision Tree	0.2564
Random Forest	0.2606
Gradient Boosting	0.2607
SVC	0.1068
Ensemble Model	0.2650

Neural Network Model

In this section, we employed a neural network model to predict the survival time of community banks. The architecture of the neural network was designed to effectively learn the complex patterns within the data, making it suitable for our prediction task.

Architecture Design and Regularization Techniques

We designed the neural network with the following architecture and regularization techniques:

1. Input Layer: Takes the standardized input features.

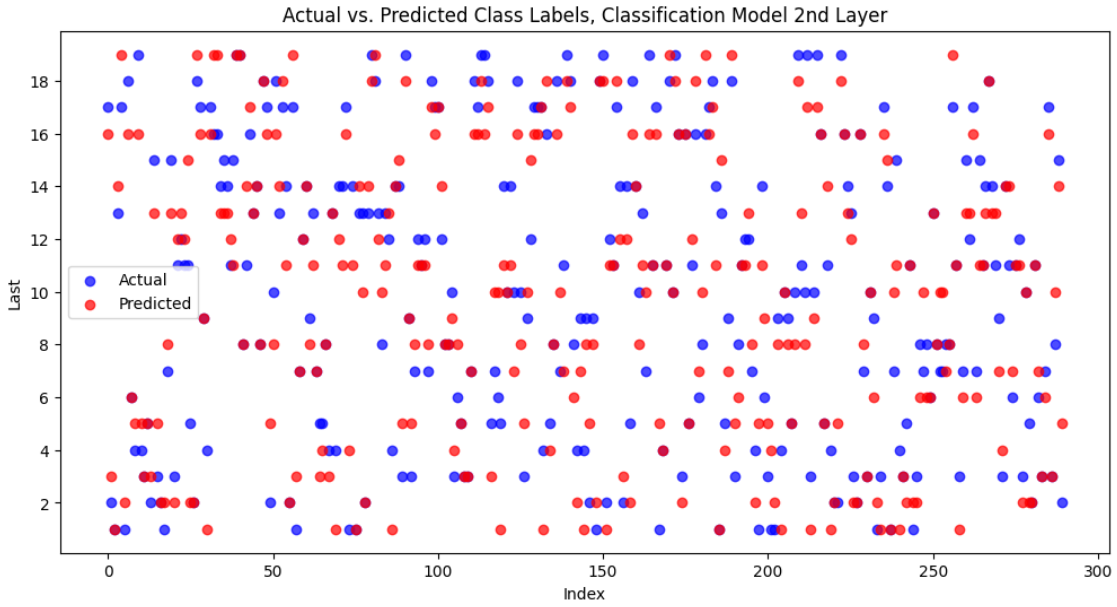

Figure 13: Ensemble Classification Model Survival Duration Prediction on Validation Set

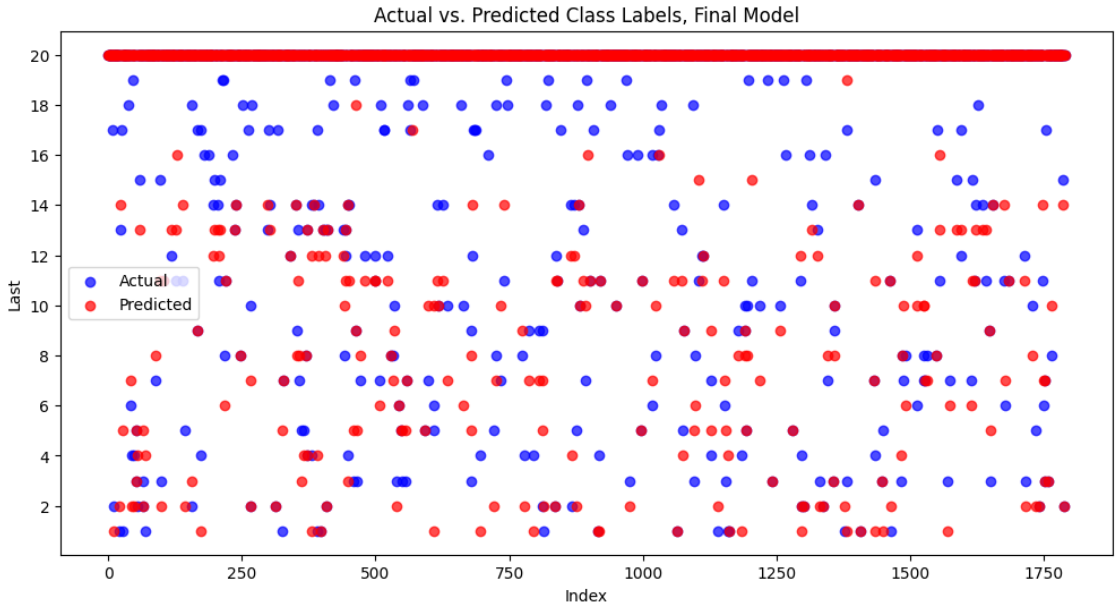

Figure 14: Final Model Survival Duration Prediction on Validation Set

2. Hidden Layers: Comprising multiple fully connected layers with ReLU activation functions and dropout layers for regularization. Additionally, we incorporated batch normalization

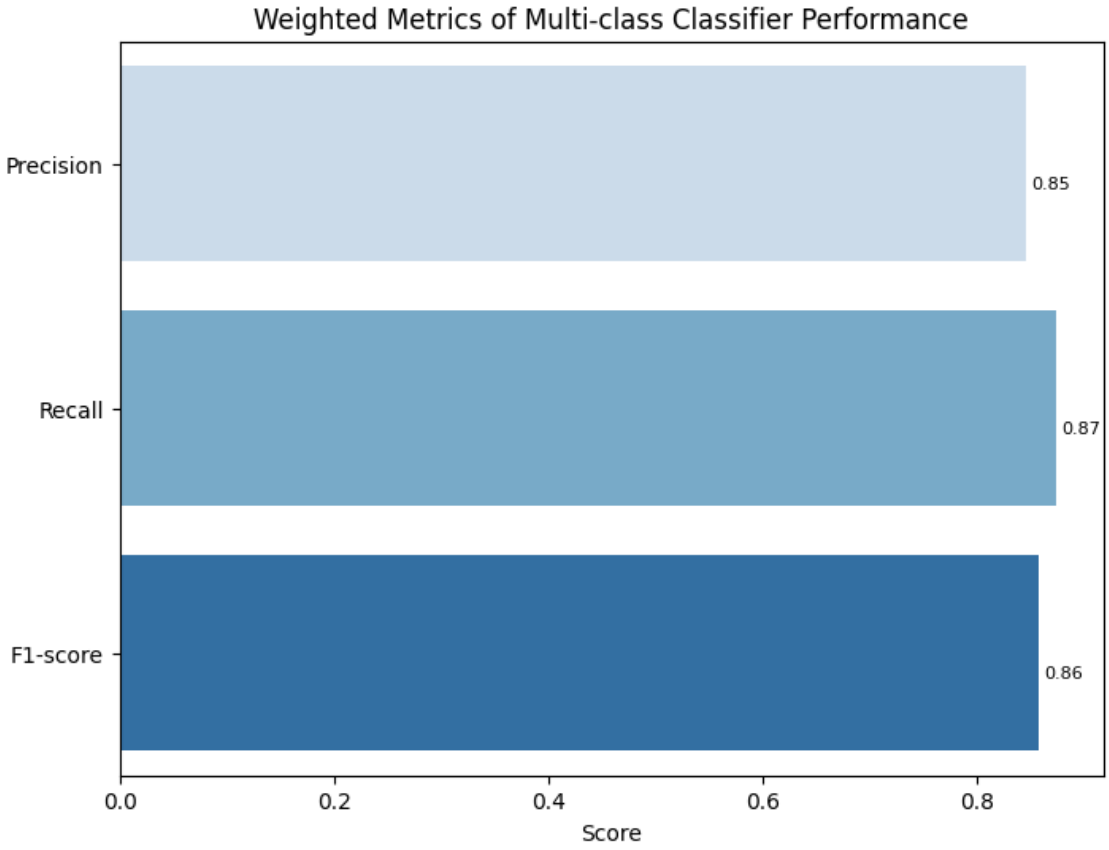

Figure 15: Model Performance Metrics

layers to improve the training speed and model stability.

3. Output Layer: A fully connected layer with the softmax activation function, producing the probability distribution over the possible survival times.

The architecture includes residual connections (skip connections) to facilitate the training process and prevent vanishing gradient issues. Residual connections allow the model to learn an identity function more easily, which in turn aids in training deeper networks.

The dropout layers are used as a regularization technique to prevent overfitting and improving generalization performance on unseen data. Dropout works by randomly dropping some neurons during training, making the model more robust by preventing it from relying too heavily on any single neuron.

Batch normalization is another technique used in our architecture. It normalizes the activations of each layer, reducing the internal covariate shift and enabling the model to train faster and more effectively.

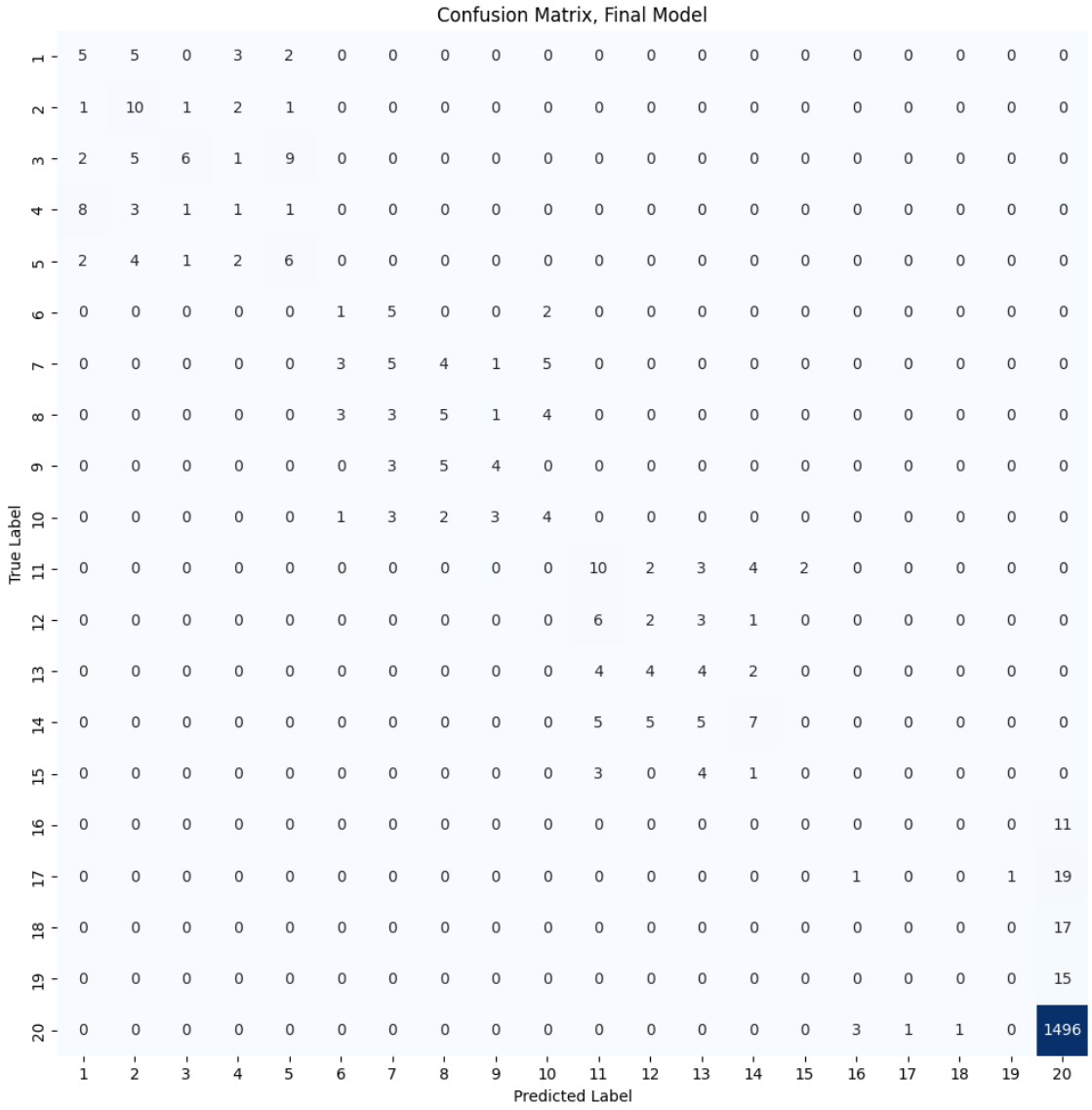

Figure 16: Final Model Confusion Matrix Survival Duration Prediction on Validation Set

Model Performance

During the training process, we monitored the performance of the neural network on both the training and validation datasets. We used accuracy as the evaluation metric, as it is a widely used and easily interpretable measure of model performance.

From the plot, we can observe that the model performance improves over time as the training progresses. However, as we increase the number of training epochs, the training accuracy

continues to improve while the validation accuracy stop increasing. This indicates that the model is overfitting the training data and losing its ability to generalize to unseen data.

To further assess the model's performance, we applied it to the test dataset, which was held out from the training process.

This plot provides a visual representation of the model's predictions compared to the true survival times, illustrating the model's ability to predict bank survival with varying degrees of success.

It is important to note that the data may not be sufficient to fully extract the real performance of this neural network architecture. Despite the use of regularization techniques such as dropout, batch normalization, and residual connections, the model continues to overfit the training data when trained for a longer duration.

Neural Network Model Summary: The purpose of our neural network model is to predict the survival time of community banks. Although the model shows potential in learning complex patterns within the data, its performance is limited by the available data. The overfitting issue indicates that additional data or further improvements in the model architecture and training process are needed to achieve better generalization performance. Nonetheless, our neural network model serves as a valuable starting point for understanding the factors influencing the survival of community banks and can be refined further as more data becomes available.

Model Prediction for Bank Call Reports in 2022

After all the modeling work, we can finally apply our final model to the bank call report data in 2022. Both the CombinedModel with two classifiers and the neural network model will be implemented to predict the survival duration for banks submitted Call Reports in 2022. Following are the model predictions for both models:

Based on the analysis presented in Figures 17 and 18, it is evident that the neural network model predicts that most banks in 2022 will endure for more than 18 quarters. Conversely, the CombinedModel suggests that most banks will only survive for five quarters, while the rest will last for more than eight quarters. This unexpected discrepancy between the predictions of the two models presents divergent outlooks for the community bank industry in the near future. A detailed analysis of this divergence is discussed in the limitations.

Project Limitations

Despite the promising results obtained using the CombinedModel and the Neural Network model, and we see several disagreements with them. For the prediction in the year 2022 using the CombinedModel, due to its limited predicting ability for longer periods, we adjust the

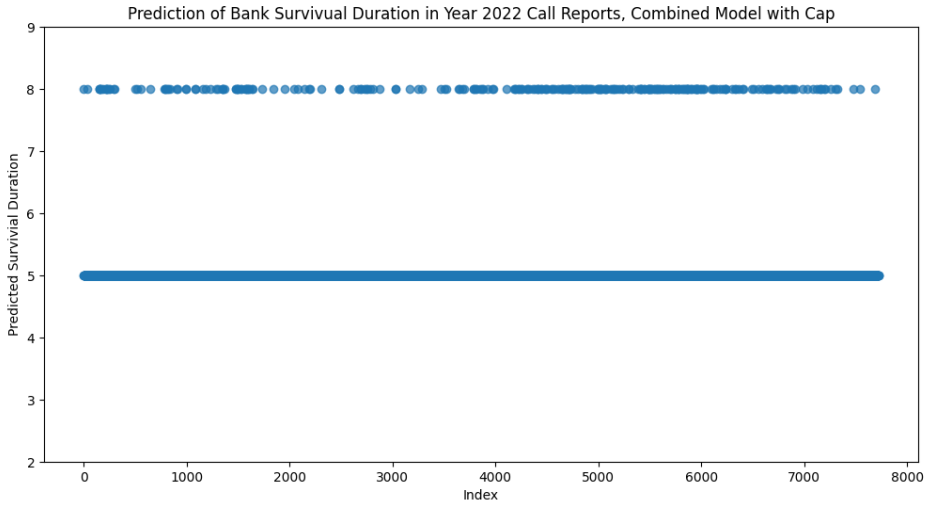

Figure 17: Combined Model Prediction on Survival Duration 2022

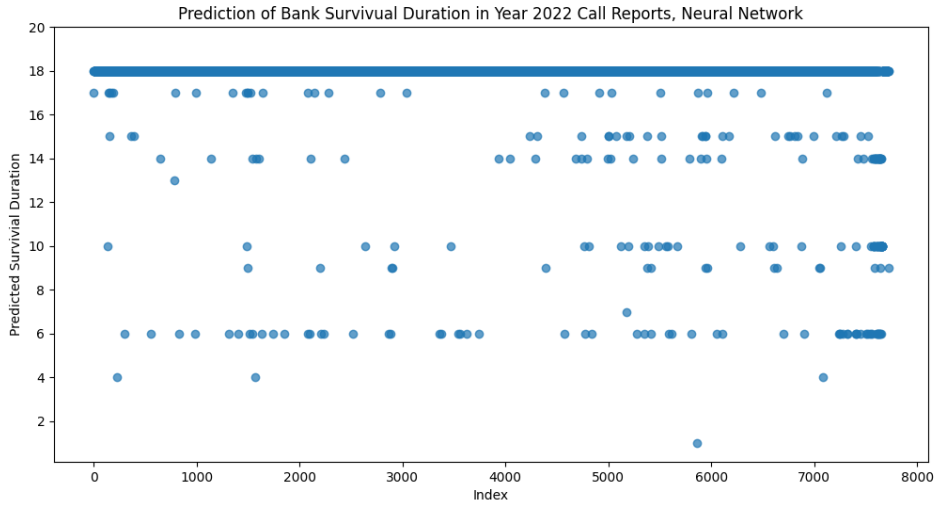

Figure 18: Neural Network Prediction on Survival Duration 2022

model prediction with a cap where if a bank's survival duration is predicted over eight quarters, we make it survives exactly as eight quarters. In addition, there are several limitations that should be acknowledged:

- Imbalanced dataset: The available dataset is unbalanced, with a majority of banks having survived the 2007-2008 financial crisis. As a result, the portion of data that can be used

to train the survival time prediction model is limited. This imbalance may lead to biased predictions and reduced performance for the minority class.

- Insufficient data for training: Due to the limited size and scope of the dataset, the models may not be able to capture the full complexity of the underlying factors that influence bank survival. The neural network model, in particular, showed signs of overfitting, indicating that additional data is needed to improve generalization performance.
- Limited Resources: Since we cannot access a higher computational power, several interesting data has not been put into the training, which may or may not help the performance of the data. The extra data include, but are not limited to: the NYSE and SP500 for the description of stock market performance; Labor market statistics; inflation rate; US Dollar Index for the international currency market, and finally, more detailed statistics of the banks.
- Data from a specific time period: The models are trained on data from the 2004 to 2008 period, which shares a similar federal funds rate as the year 2022. However, we do not use the 2022 dataset for training, as we cannot create the labels for this data. This limitation may affect the model's ability to generalize to different economic environments and financial conditions.

Besides, since we assume the bank's report better summarizes the impact to them given by the federal funds rate, we did not include it as a feature. Though it may help increase the performance of our train test, it may suffer from over-fitting, unstable estimates, and inaccurate errors due to the risk of bringing extra collinearity.

Conclusion

In conclusion, this project has examined the survival of community banks in a rapidly rising interest rate environment, utilizing various machine learning techniques such as regressions, multi-classifiers, and neural networks. Although our models achieved a commendable accuracy rate of over 85% when predicting the survivability of community banks during the period of 2004-2008, the accuracy of predicting survival time was found to be relatively low.

Notably, using the CombinedModel and neural network to predict community bank survival rates produced divergent results, which may be attributed to the high collinearity among three variables, leading to sub-optimal model performance.

Moving forward, several ways to enhance our model performance include collecting additional data points, refining feature selection, or exploring more advanced modeling techniques. Addressing these limitations will further our understanding of the survivability of community banks in a dynamic interest rate environment and provide valuable insights for relevant stakeholders in the banking industry.

1. Collecting more data for time series prediction on Long Short-Term Memory (LSTM) networks. This method is capable of learning long-term dependencies, especially in sequence prediction. In our case, 20-time points cannot support a time series prediction.
2. Using LMI

This project aims to offer insights into the performance of community banks by analyzing the Call Reports submitted by each bank at the end of every quarter. Our model outputs are intended to serve as a reference for evaluating a bank's performance during rapidly rising interest rates, facilitating informed financial decisions for banking professionals and the general public.

To this end, we have developed a dashboard based on our model, which provides an intuitive and user-friendly interface for accessing key performance metrics of community banks. This dashboard has the potential to enhance transparency and promote data-driven decision-making within the banking industry.

References

[1] Community banking research program - study reference data. Online.

[2] FFIEC. Download bulk data - ffiec central data repository's public data distribution. Online, 2023. Accessed: April 23, 2023.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 2013.