Time is the Witness: Bank Failure Prediction via a Multi-Stage Model with Artificial Intelligence

Dimitrios Gounopoulos[*1], Emmanouil Platanakis[†1], Haoran Wu[‡1], Wenke Zhang [§1]

[1]University of Bath, School of Management

Abstract

Bank failure prediction is a popular topic that requires highly accurate results. We contribute to the literature by determining whether models based on the crisis data are suitable for predicting bank failure during a stable period and which predictors can be held for long-term forecasting. In this paper, we design a multistage procedure, including feature selection and application of four advanced single machine learning techniques to predict bank insolvencies based on a sample of U.S. banks over 2006-2019. We set two time windows to explain the bank failure in pre and in-crisis and predict bank insolvency in a stable period. The feature selection results illustrate that capital, asset, and liquidity predictors contribute more to explaining bank failure. In the prediction, we found that the second-generation ensemble method has a superior prediction performance and provides the most accurate prediction results. We also extend the multi-step predicting approach to reclassify banks into four and six groups, which takes an insight into banks' risk default levels and compares banks' risk-bearing abilities in different risk levels. Our results of the multigroup classification suggest that non-failed banks can have more risk than failed banks. With adequate liquidity and capital, the non-failed banks can bear more risks, which helps them survive during the crisis.

[*] D.Gounopoulos@bath.ac.uk

[†] E.Platanakis@bath.ac.uk

[‡] hw2258@bath.ac.uk

[§] wz798@bath.ac.uk

1 Introduction

Banks play an essential role in financing the economy as intermediaries in financial markets. Due to the contagion effects, the bank failure triggers much more severe volatility than the failure of other business firms. According to the Federal Deposit Insurance Corporation, 24 banks failed in the U.S. from 2000-2006, while the number increased to 492 during the 2009 financial crisis. The increasing complexity and volatility of financial markets lead to the need to identify the potential risks in banking systems. The implementation of Basel III emphasizes the critical role of early regulation. At the same time, increasing risk management concerns the rise of bank failure prediction topics with various indicators and techniques.

Many existing studies (Sarkar and Sriram, 2001; Ramirez and Shively, 2012; Berger, 2013; Shaban and James, 2018; Boyallian and Ruiz-Verdú, 2018; Mili, Khayati and Khouaja, 2019) investigate bank financial performance regarding their risk-taking behavior. The motivation for predicting bank failure is to identify banks that are at risk of collapsing and consequently take action to prevent it. This can help protect depositors, investors, and the broader economy from the negative consequences of a bank failure, such as loss of access to funds, increased financial instability, and potential economic downturn. In the short run, early warning systems for bank failure can also help regulators and policymakers identify and address underlying issues within the financial system that may be contributing to a bank's struggles. In the long run, accurate predictions can lower the costs brought by the crisis and help to stabilize the financial markets. Moreover, accurate results of a possible bank failure can considerably help supervisors adjust regulations further.

Banks generate vast amounts of data, including financial statements, lending and investment activity, and customer behavior, which can be challenging to process and analyze in a timely and accurate manner using conventional manual methods. Artificial intelligence (AI) and machine learning (ML) are beneficial for predicting bank failures because they can analyze large volumes of complex data and identify patterns and trends that may not be able to discern in analysis through personal efforts. On the other hand, AI and ML algorithms can analyze large amounts of complex data quickly and efficiently since they can learn from data in real time and identify trends that are not apparent to humans. Therefore, AI and ML are suitable for identifying potential indicators of bank failure, such as declining asset quality, reduced liquidity, and increased leverage, when they occur rather than after the fact. This is particularly useful in predicting bank failures, where prompt and accurate analysis is crucial to decision-making, helps provide early warning of potential problems, and allows banks and regulators to act before disruptions occur. Additionally, AI and ML algorithms can be continuously trained and updated with new data, making them more accurate and effective.

Regarding the mentioned outstanding benefits of using AI and ML, a strand of existing research employs ML techniques to improve the classification accuracy in predicting bank failures. Some use stand-alone methods, such as boosting methods(Carmona, Climent and Momparler, 2019; Climent, Momparler and Carmona, 2019), support vector machine (SVM) (Ecer, 2013; Manthoulis et al., 2020), Neural Network (NN) (Tam and Kiang, 1992; Ng, Quek and Jiang, 2008). Besides, some research employs various machine learning methods to test the efficiency of each technique and find the best-performing one(Ecer, 2013; Erdal and Ekinci, 2013; Uthayakumar et al., 2020). Thus, they are aimed to reach higher accuracy in forecasting. Additionally, the explanation power of predictors also plays an essential role in forecasting accuracy (Cole and Wu, 2014). A group of studies investigates the influence of macroeconomic factors (Apergis and Payne, 2013; Mare, 2015; Wulandari and Kusairi, 2017), aand some other studies focus on some individual factors of banks (Berger, Imbierowicz and Rauch, 2016; Boyallian and Ruiz-Verdú, 2018).

We contribute to the literature by focusing on whether models based on the crisis data are suitable for further predicting bank failure in a stable period and which predictors can be held for long-term forecasting. Using a comprehensive dataset incorporating bank failures from 2006 to 2019 in the U.S., we employ a multi-stage approach to explain and predict bank failures over time. The output in the last stage will be used as the input in the successive stages; for instance, the optimal results of the feature selection will be devoted to the next stage of bank insolvency prediction to explain the crisis of failure caused by poor operations within banks. The motivation behind using a multi-stage approach is to make the process of bank failure prediction easier to understand and solve by breaking it down into smaller, more manageable parts. This research breaks the predicting process into three portions: correlation test, feature selection, and two-group and multi-group classification. In each step, we unit the results of different methods.

Additionally, a multi-stage approach allows for a more structured and methodical approach to problem-solving, which helps ensure that all relevant factors are considered and the classifications are efficient. For instance, our feature selection procedure encompasses the evaluation of correlation, non-linear relationships, and multicollinearity. Finally, the multi-stage approach is more flexible and robust than the single methods to changes in data, as it allows the combination of strengths of different methods. Single traditional methods can be sensitive to data type, quality, and endogenous problems. At the same time, a multi-stage approach is tailored to solve all potential problems most efficiently.

The view that a financial institution has a culture or business model that affects its sensitivity to crises implies that the performance of a financial institution in one crisis should predict its performance in another crisis (Fahlenbrach, Prilmeier and Stulz, 2012). If it is the case, explanations of bank performance during the financial crisis can provide a measure of its exposure to failure over time. In this case, we set two time windows, 2006-2012 and 2013-2019. The former illustrates the bank's failure before and during the financial crisis. The latter time window covers the post-crisis period to the stable recovery, which shows the bank's performance post the financial crisis. The second time window presents the persistence in a bank's risk culture, making it more sensitive to exposure to failure, even in a thriving financial market. We use the data for training the ML models and provide insights into how bank failure occurs.

In the multi-stage procedure, we implement feature selection in the first and second stages, which include the correlation test and embedded methods. It is noted that discovering the relevant predictive variables is an essential preprocessing to solve the optimization problem, such as accurately predicting bankruptcy, especially when one is faced with a vast data set including over one hundred thousand pieces of information. Also, variable selection is crucial when the true underlying model can show sparse results to identify the remaining significant predictors (Zou, 2006). In general, feature selection is a process to exploit specific feature criterion that determines the salient features that are relevant but not redundant to the classification tasks (Unler and Murat, 2010). Thus, identifying the significant features guarantees efficient prediction performance in applied models and highly accurate results(Maldonado, Pérez and Bravo, 2017).

A strand of existing research used individual feature selection methods to indicate the performance improvement in regression (classification) (Unler and Murat, 2010; Maldonado, Pérez and Bravo, 2017; Ghaddar and Naoum-Sawaya, 2018; Petropoulos et al., 2020). However, some potential challenges are still in the way of efficiently applying individual feature selection methods. The research to date has tended to focus on continuous data rather than categorical data. Our paper predicts bank failures by classifying banks into failed or non-failed groups based on the selected features, resulting in a binary dependent variable. In this case, a simple correlation test, such as Spearman's correlation, is unsuitable for measuring associations between predictors and bank failures.

Furthermore, appropriate feature selection methods for such categorical data are limited due to the difficulty in capturing the internal link between variables that are only significant if they associate with some meaningful values. Thus, some widely used penalized methods, such as Lasso, cannot select the best variable subset when applied to the categorical data because such methods require continuous variables in linear regression. Therefore, we must apply an improved penalized regression model by combining the likelihood function. The penalty term is used to regularize a logistic regression model for classification tasks with categorical variables.

In terms of computational efficiency, the individual feature selection method is exposed to limitations of computational consumption and high variability(Unler and Murat, 2010). One of the critical criteria for feature selection is a benchmark that evaluates each feature's significance and determines the subset groups' size. However, the benchmark setting is subjective to users and varies across methods, which leads to different explanations of feature importance. Thus, the salient feature identified by one method may not be recognized as meaningful by another. Besides, in terms of the classification problem, lacking information about the interaction between the features and the classifier also lead to variability across subsets(Bennasar, Hicks and Setchi, 2015). Additionally, the non-linear relationship between the dependent and independent variables is another challenge in generating efficient results when the explanation power of predictors is overestimated. Past studies on bank distress either result in various explanatory variables in the form of financial ratios constructed by domain knowledge or solely focus on data mining that reduces the volume of predictors to improve the prediction. In these cases, the explanation to ascertain the association within the vulnerability indicators is far from the complete story. Identifying the non-linearity and applying suitable models are critical for investigating the internal causation of synchronous change of variables. Otherwise, the ineffective model may retain redundant and irrelevant features, resulting in an overestimated feature significance.

To solve the abovementioned issues, the correlation test and embedded methods in the first two stages are combined to improve the performance of the prediction. Using a multi-stage approach will be made more explicit by considering the properties in the feature selection models when the dependent variable is a dummy. In the first stage, we used mutual information selection that uses entropy to evaluate the information gained from one variable based on the knowledge of the other to determine the degree of correlation between variables. In the second stage, Logistic Lasso and the Random Forests (RFs) further shrink the volume of variables. The Logistic Lasso is an improvement on the traditional Lasso that uses logistic regression as the fundamental function when dealing with classification problems involving categorical dependent variables. Random Forests (RFs) is an ensemble tree method suitable for a non-linear relationship evaluation of the feature importance. Consequently, the optimal variable subset is the combination of results from Logistic Lasso and RFs. Specifically, most variables in the optimal subset are derived from intersections of the results provided by both embedded methods. After accounting for non-linear correlations, the RFs discover the rest factors that are important but not revealed by the Logistic Lasso. Thus, the ensemble technique units the benefits from each method and diversifies the errors by considering the data type, non-linear relationship, and variability.

In the third stage, we predict bank insolvency primarily by using different types of ML models that are popular and widely accepted. To be more objective, we will compare their performance regarding predicting accuracy and identify the most efficient one for long-term forecasting. We apply supervised machine learning models, including the type of boosting (i.e., gradient boosting decision tree, GBDT), deep learning (i.e., neural network, NN), hyperplane classification (i.e., support vector machine, SVM) as well as a distance-based technique, the K-nearest neighbor (KNN), to non-linear data. These methods are extensively accepted by the studies of differentiating failure from solvent institutions, while few emphasize comparing their characteristics. Moreover, GBDT is barely used in bank failure prediction.

Our objective is not simply to explain the previous failure but, more importantly, to predict the failure in the future. Therefore, we focus on prediction accuracy improvement and investigate alleviating different errors, such as reducing the impact of noise data and redundant information. Thus, we use four single models in the third stage to diversify the application of the bank failure prediction and provide comprehensive evaluations of the bank performance based on the selected variable set. In the meantime, we would like to identify the most efficient machine learning model that can have the best performance for forecasting extensive banking data.

Additionally, banks risk default if they are impeded from meeting their financial obligations to depositors and creditors. Therefore, it is vital to understand whether some banks are prone to perform poorly during the financial crisis related to their capability to bear the risks. In some related papers (Fahlenbrach, Prilmeier and Stulz, 2012; Beltratti and Stulz, 2012; K"ohler, 2015), the bank's business model and risk culture are persistent in explaining its performance during the crisis. Our research also focuses on the risk culture that presents by banks' risk-taking behavior. To answer the question on the relationship between a bank's poor behavior and its capability of bearing risks, we calculate the bank's risk level, z-score, and use it as an indicator, along with banks' solvency, to reclassify observed banks into multiple groups. Our finding suggests that banks' ability to deal with risks is a more significant factor in explaining a bank's crisis performance than its absolute risk levels. Banks that can effectively resist negative shocks have a higher probability of survival, even at higher risk levels.

Our main results hold up in a variety of robustness tests. The in-sample results indicate GBDT outperforms in classification. Specifically, It has the highest rates in all performance measurements. Compared to the other three models, GBDT has superior performance because it can capture non-linear interactions between features and is robust to outliers and noise in the data. Furthermore, GBDT uses the boosting procedure to continuously improve its performance by focusing on the portions of the feature space with poor performance.

KNN performs better on the "in-sample" and "out-of-time" samples.

Moreover, NNs have the worst predictive power across all the samples. For the multi-group classification, the best performance of GBDT shows evidence of specific default risks within and between the failed and non-failed banks. The minor decline in prediction accuracy from a four-group classification to a six-group classification is the consequence of the decrease in the data distribution per group as the number of groups grows.

The rest of the content proceeds as follows: Section 2 reviews the previous studies on bank failure prediction. Section 3 describes the data of banks. Section 4 is devoted to the methodologies of variable selection and prediction models. Section 5 presents the prediction results. Section 6 summarizes and concludes our study and introduces some further research plans.

2 Literature Review

The popularity of machine learning models attributes to the development of computing power and their ability to learn patterns in data and make a prediction without being explicitly programmed, which allows them to be applied to many different fields, including finance, healthcare, and marketing. A strand of existing literature conducts comprehensive surveys of bank failure prediction with machine learning methods. Fethi and Pasiouras (2010) focused on bank performance evaluation, such as bank insolvency forecasting, using machine learning methods. Demyanyk and Hasan (2010) provided an inclusive summary of the machine learning methods used for elaborating, predicting, and taking the remedial actions the bank defaults. In recent research, Manthoulis et al. (2021) and Doumpos et al. (2022) presented the comparative results of the summarized machine-learning approaches to bank failure prediction with different prediction horizons and variable sets. The existing literature has illustrated that the improvement of bank failure prediction also demands decent predictors for accurately explaining bank insolvency with the help of machine learning.

2.1 Selected Features

A large group of studies in bank failure prediction utilized CAMELS (capital, asset, management, equity, liquidity, sensitivity to risks) as predictors (Curry and Shibut, 2000; Barth, Trimbath and Yago, 2006; Cole and White, 2012; Mayes and Stremmel, 2012; Vazquez and Federico, 2015; Audrino, Kostrov and Ortega, 2019; Manthoulis et al., 2020). Moreover, some research combined CAMELS with additional variables to enhance the explanatory power, which can help increase the prediction models' accuracy (Serrano-Cinca et al., 2014; Manthoulis et al., 2020). Besides CAMELS, a group of literature used study-specific indicators to explain the bank failure, such as audit quality (Jin, Kanagaretnam and Lobo, 2011), the bank's internal control (Jin et al., 2013), the role of bank ownership, management and compensation structure (Berger, Imbierowicz and Rauch, 2016), local housing conditions and internal bank funding (Sun, Wu and Zhao, 2018), the lag effect of bank efficiency (Assaf et al., 2019), cost of insured deposits (Chernykh and Kotomin, 2022). Using predictors beyond the CAMELS increases the diversification in the variable sets.

Furthermore, variable selection methods are applied to provide conclusive evidence of the importance of the variables in predicting bank insolvency. There are two domain ways, feature selection, and dimensionally reduction—the former selects features without changing them, such as Lasso (Tibshirani, 1996). The latter is to transform the features into a lower dimension, such as Principle Component Analysis (PCA). Before forecasting, some studies use the single feature selection method (Canbas, Cabuk and Kilic, 2005; Carmona, Climent and Momparler, 2019). A few pieces of literature apply multiple-stage variable selection. Petropoulos et al. (2020) select the best variable subset by combining filter methods, including Pearson correlation and pairwise tests, and Embedded procedures, including Lasso.

This research uses CAMELS as primary predictors that provide comprehensive financial information on banks. CAMELS are widely used in banking studies because they are objectively evaluated based on clearly defined criteria, which means they are the same for all banks and are not subject to personal interpretation. Therefore, CAMELS are better than individually determined factors in terms of consistently using and assessing the financial health of banks in an unbiased manner. In the feature selection process, we consider the correlation between predictors and the binary dependent variable and the non-linear relationship and multicollinearities among the predictors. These issues still need to be fully addressed in the feature selection process in the existing literature. We use a multi-stage approach to refine the feature selection, including testing correlation. We eliminate less relevant features by adding a penalty term to the loss function and ranking the feature's importance. Our contribution is to identify the characteristics with solid explanatory power for bank failure and demonstrate that they can be used to predict poor bank performance over time.

2.2 Development of Prediction Methods

At the same time, bank failure prediction methods range from conventional methods, such as logit regression, to advanced machine learning techniques. Huang, Chang, and Liu (2012) use the logistic regression model for bank failure prediction for developing and developed country cases. Similarly, Serrano et al. (2014) apply the partial least square path modeling (PLS-PM) and logistic regression model to the U.S. data to investigate the pre-condition of the 2009 banking crisis. Cleary and Hebb (2016) utilized the discrimination analysis to the in-sample data in 2002-2009 and successfully distinguished solvent banks from insolvent banks.

Some studies indicate that Neural Networks have the best performance for bank failure prediction. In the early research, Tam and Kiang (1992) discussed that even though NNs are criticized for being time-consuming and difficult interpretation, their performance surpasses other prediction models. In the following study, Boyacioglu and Kara (2009) indicated that NN-related techniques, multi-layer perceptron, and learning vector quantization have the best predictive power (100%) on training and validation data. Moreover, some research notes the superior performance of the SVM. Gogas, Papadimitriou and Agrapetidou (2018) pointed out that, with the statistically selected explanatory variables, the accuracy of SVM classifying the solvent and insolvent banks reaches 98% and 99.22% for in-sample and out-of-sample, respectively.

In recent literature, many advanced machine-learning techniques and extensions have been adopted for bankruptcy prediction. Carmona, Climent and Momparler (2019) evaluated the predictive power of XGBoost, the extended branch of boosting methods, and compared it with conventional models. The results show that XGBoost is the most efficient method for bank failure prediction based on the selected sample. In line with Carmona, Climent and Momparler (2019), Pham and Ho (2021) also reported that XGBoost has higher predictive power than AdaBoost and Gradient Boosting in forecasting bank insolvency.

However, the number of studies using hybrid machine learning models is limited in bank failure prediction. Ekinci and Erdal (2017) compared the hybrid machine learning models based on trees with the single models for tackling the bank failure problem and found that the former yields superior performance. Petropoulos et al. (2020) examined the ability of SVM, NN, and RF and some conventional models, including logistic regression and linear discriminant analysis, to forecast bank insolvency based on the in-sample and out-of-sample data. The results show the superior prediction ability of RF in both the U.S. case and the European case.

Our study aims to examine the consistency of prediction indicators by analyzing the pre- and post-crisis periods. The related paper, Goenner (2020), shows that policymakers using estimates based on the Savings and Loans crisis would identify the banks in critical condition and unhealthy in early 2009. Our study differs from existing literature in that we argue that the financial conditions influencing bank failure in the 2009 financial crisis also affected failure in the stable period. Our research aims to show that the same predictors used to explain poor bank performance during the 2008 financial crisis can predict bank performance and the probability of failure in the post-crisis period. Moreover, we utilize four single machine learning methods and compare their performance to identify the most suitable one with superior prediction power. The comparison illustrates that no "one size fits all" method exists, but improving the ML model will lead to increased prediction accuracy. We also reclassify banks into multiple groups and take an insight into the influence of risk default level on bank failure, which compares banks' risk-bearing abilities during the crisis.

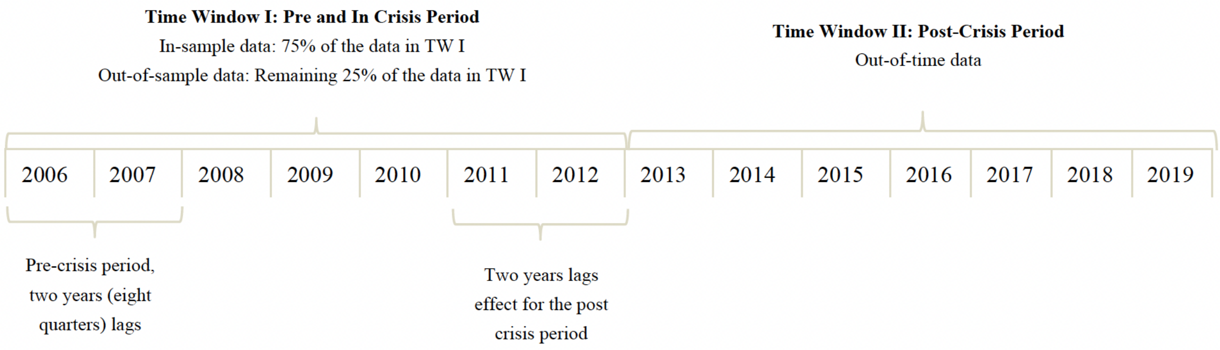
3 Data Description

The bank failure predictors are collected from the Federal Financial Institutions Examination Council (FFIEC). Most financial characteristics are obtained from Reports of Condition and Income (Call Reports) and the Uniform Bank Performance Reports (UBPRs) in FFIEC.

Specifically, the Call Reports contain bank income statements, balance sheets, loan information, and other information reflecting bank health. At the same time, the UBPRs show the impact of management decisions and economic conditions on a bank's performance and balance-sheet composition to evaluate the adequacy of earnings, liquidity, capital, asset and liability management, and growth management. There are 8020 banks covered over the sample period 2006-2019, and the bank information is classified quarterly.

To evaluate the bank performance across time and samples with the same explanatory variables, we split the data into two parts by time. In the first part, the data are resampled to two groups based on the rule-of-thumb, the in-sample data and the out-of-sample data, where the former is randomly stratified of 75% of 2006-2012 data for training, and the latter comprises the remaining 25% of the observations of 2006-2012 for testing. The first part illustrates the bank performance before and during the financial crisis, which composes the basis of selecting the authentic explanatory features. The second part indicates the out-of-time sample that discovers the bank performance in the stable period spanning 2013-2019. Figure 1 presents the two time windows used in this study.

Figure 1: The Time Window of Data Groups


Time Window I: Pre and In Crisis Period
In-sample data: 75% of the data in TW I
Out-of-sample data: Remaining 25% of the data in TW I
Time Window II: Post-Crisis Period
Out-of-time data
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Pre-crisis period, two years (eight quarters) lags
Two years lags effect for the post crisis period

The information about failed banks is collected from the Federal Deposits Insurance Corporation (FDIC), which lists bank failure events. In the period 2006-2019, the sample contains 535 failure events. Table 1 shows the sample data of failed banks during this period.

Table 1: The Number of Failed Banks in Year 2006-2019

Time	Failed banks
2006	0
2007	3
2008	25
2009	140
2010	157
2011	92
2012	51
2013	24
2014	18
2015	8
2016	5
2017	8
2018	0
2019	4

The variable selection is based on the CAMELS (capital, asset, management efficiency, earnings, liquidity, and market risks) that are both globally accepted by economic authorities and extensively applied in academic research (Lopez, 1999; Bank and Fund, 2005; Audrino, Kostrov and Ortega, 2019). FFIEC and FDIC provide comprehensive information about CAMELS of both failed and non-failed banks. Table A1 shows variable descriptions in detail, and A2 shows the summary of variables by groups (failed and non-failed). Besides CAMELS, we include the bank size, defined by the logarithm of the assets, as another variable labeled by O1. Some studies (i.e., Bertay, Demirg"uç-Kunt and Huizinga (2013)) on risk predictions noted that banks' incentive to chase risks is positively associated with banks' size as bigger banks can take the privilege of ”too big to fail”.

Moreover, Zhao, Sinha and Ge (2009) identified that using feature construction data is better than using absolute values in the regression models. Some features are less relative in describing the characters individually but can become relevant when combined (Markovitch and Rosenstein, 2002). Thus, this research uses absolute values and feature construction data to diversify bank performance indicators. Moreover, we use two-year (eight quarters) lags before the financial crisis in 2008. Unlike the existing studies that use one year lag, our expanded time window to increase the data volume for evaluating the causal effect of bank performance before the crisis on the following bank failure (Cole and White, 2012; Audrino, Kostrov and Ortega, 2019; Manthoulis et al., 2020; Petropoulos et al., 2020). Ultimately, The base for variable selection consists of 405 predictors, including original explanatory and lagged-explanatory variables.

6801 banks compose the sample set without missing values. Meanwhile, the independent variable is the dummy variable that equals one if the bank failed and zero otherwise. One common problem is that the number of non-failed banks is overwhelmingly larger than that of failed banks. In terms of imbalanced data, oversampling such as Synthetic Minority Over-sampling Technique (SMOTE) is widely accepted to rebalance the data set by adding estimated data to the minority group. However, if the imbalanced gap is overwhelmingly large, estimated data cannot truly represent the actual data due to outliers and noises. Besides, many estimated data complemented to the minority group will lead to the overfitting problem. Thus, following Petropoulos et al. (2020), which employs the under-sampling method, we narrow the imbalance in the training sample by randomly selecting 10% from the majority group, and the proportion of the minority to the majority increases from 3.4% to 40%. Compared to the existing literature, We do not deal with the remaining testing data because any estimation action in the testing groups will lead to overfitted prediction results. Moreover, our feature selection and bank failure prediction results are convincing because we have a comprehensive set of variables that captures the performance of banks.

4 Methodology

This section introduces the feature selection and classification methods in the multi-step approach, including the correlation test, two embedded feature selection methods and four machine learning classification methods. The implemented procedure is presented in the flow chart (figure A1). We used the combined results of three feature selection methods to propose a sound variable subset employed for prediction in the next stage. The following subsections thoroughly clarify the details and discussions of the methodologies in each phase.

4.1 Variable Selection

We used three different types of feature selection models, Mutual Information Selection, Logistic Lasso and Random Forests, that are combined to generate the best variable subsets. They have different mechanisms with the same goal of selecting a compact set of superior features at very low cost. Specifically, the Mutual Information Selection filters the redundant and irrelevant information by detecting dependencies between the independent variables and the dependent variable. The Logistic Lasso produces sparse results by shrinking the parameter of less relevant variables to zero. Also, the logistic lasso is an automated feature selection procedure and overcomes the challenge of multicollinearity. The Random Forests is a non-linear feature selection method that shows the feature importance based on the mean decrease in impurity. The combination of three methods increases the diversification that eliminates the influence of errors. Therefore, it is evident that selected variables are relevant to the dependent variable in great dependencies and feature importance. The following content elaborates on each method in detail.

4.1.1 Mutual Information Selection (MI)

Mutual information (MI) selection (Shannon, 1948) is a filter method that measures the mutual dependencies between two variables. It indicates how much information can be obtained from a random variable based on the knowledge of another random variable. In this research, we are proceeding with the binary classification, therefore, it is important to accurately measure the correlation between the categorical variable and the continuous variable. MI is suitable for correlation tests without restriction on data type.

The entropy is used as the basic criteria for measuring mutual information. Based on the information theory of Shannon (1948), entropy measures uncertainty related to the probability of occurrence of an event. Formally, for one random variable x with a possible value

$\{x_1, x_2, \cdots x_n\}$, entropy is defined as:

$$H(x) = -\sum_{i=1}^{n} p\left(x_i\right) \log\left(p\left(x_i\right)\right) \tag{1}$$

where $p\left(x_i\right) = \Pr\left\{x = x_i\right\}$ is the mass probability of random variables. The entropy can be interpreted as the negative expected value of the logarithm of the mass probability. For two random variables x and y that have joint mass probability $p\left(x_i, y_i\right)$, the joint entropy is as follows:

$$H(\{x, y\}) = -\sum_{i=1}^{n} \sum_{j=1}^{n} p\left(x_i, y_j\right) \cdot \log\left(p\left(x_i, y_j\right)\right. \tag{2}$$

It measures the uncertainty contained by both variables. The remaining uncertainty of one random variable x when the value of y is given is measured by the conditional entropy that is defined as:

$$H(x \mid y) = \sum_{j=1}^{n} p\left(y_j\right) \cdot H\left(x \mid y = y_j\right) \tag{3}$$

With the basic concepts of entropy, MI is defined as the follows:

$$I(x; y) = \sum_{i=1}^{n} \sum_{j=1}^{n} p\left(x_i, y_j\right) \cdot \log\left(\frac{p\left(x_i, y_j\right)}{p\left(x_i\right) \cdot p\left(y_j\right)}\right) \tag{4}$$

MI is bigger if two variables are highly correlated, while MI becomes zero if two variables are statistically independent. For better interpretation, MI can be expressed as:

$$I(x; y) = H(y) - H(y \mid x) \tag{5}$$

Equation (5) shows that the correlation between two random variables depends on the difference between the entropy of one variable and the conditional entropy of the other one, which indicates the reduction of uncertainty on the value of one random variable once the other one is known. MI identifies the union information of two variables to measure their dependencies and correlations. As $H(y)$ is independent, the feature selection in order to maximize the $I(x;y)$ becomes the work of minimizing the conditional criteria $H(y \mid x)$. Moreover, the MI is favorable for three properties: first, it has the capacity of measuring the correlation between any kinds of variables, such as categorical and continuous one; In our research, we have dummy dependent variable , where other linear models are not able to capture the correlations within our variables. Second, MI has the advantage of detecting non-linear relationships between variables, which is the best suitable for our categorical data; Third, it is invariant under space transformation, which means the logarithm of our data do not influence the correlation results when using MI.

4.1.2 Embedded Methods

Logistic Lasso It is acknowledged that Lasso is a widely used feature selection method (Meier, Van De Geer and B"uhlmann, 2008; Rapach, Strauss and Zhou, 2013) It has advantages in highly efficient performance and overcomes the multicollinearity in regression. However, Lasso becomes disgraced for the classification problem in which labels of given responses are discrete values $(i.e., 0 or 1)$ as it is incapable of dealing with the regression with the categorical dependent variable. Thus, in this research, we employed the combination of Logistic regression and Lasso, where the former helps the latter to deal with the regression problem based on the binary dependent variable. In linear regression with continuous variables, the intuition behind Lasso is to minimize the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant(Tibshirani, 1996). In the Logistic Lasso, the penalized element is preserved from Lasso. For feature selection in the classification problem, the Logistic Lasso uses the likelihood function, and the parameter estimates are obtained by maximizing the log-likelihood function with l_1 penalty. Formally, the process of achieving the Logistic Lasso is as follows: For x_i and corresponding response

$y_i \, (i = 1, 2 \ldots n)$, the likelihood function is

$$P\left(y_i = 1\right) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \tag{6}$$

In the logistic regression, the estimated parameter is obtained by maximizing the log-likelihood function:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \left[y_i \log \left(\frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right) + (1 - y_i) \log \left(1 - \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right) \right] \\ &= \sum_{i=1}^{n} \left[y_i x_i \beta - \log \left(1 + e^{x_i \beta} \right) \right] \end{aligned} \tag{7}$$

The Logistic Lasso is the maximization of penalized log-likelihood function that achieves the regularization by integrating the regression of binary variables and the coefficient shrinking. Formally, it is defined as following:

$$l(\beta) = \sum_{i=1}^{n} \left[y_i x_i \beta - \log \left(1 + e^{x_i \beta} \right) \right] - \lambda \sum |\beta| \tag{8}$$

The logistic Lasso unions the advantages of Logistic regression and Lasso, but also circumvents the shortcomings of both methods. Specifically, it is better than the original Lasso to solve the classification problem with the linear regression, and suitable for our research that needs the likelihood function to measure the categorical information of the independent variable. It also performs superiorly to Logistic regression by trading off a small increase in bias for a large decrease in the variance of the predictions.

Random Forests (RFs) RFs start by drawing a bootstrap sample from training data. Then, it grows a random tree by selecting a group of variables randomly. The tree grows from the root node to the leaf nodes, with intermediates called internal nodes. RFs pick the selected group's best variable and split the node into two sub-nodes in the procedure. Then, recursively repeat the selection and splitting until reaching the minimum node size.

The main advantage of RFs is that they can capture complex data interactions and have a relatively low bias if the tree grows efficiently large. The RFs output is the class selected by most trees for classification. Thus, RFs are less prone to overfitting and generally perform better than simple decision trees.

The feature importance of the RFs is evaluated using Gini importance, which determines the split at each node. It is calculated as the decrease in node impurity weighted by the probability of reaching the node. The probability is measured by the proportion of samples reaching that node to the total number of samples. The higher feature importance indicates higher explanation power and relevance. Thus, the RFs select the features in the classification procedure and group the subsets of features with high purity.

4.2 Classification Methods

We employ four popular single machine learning classification methods in the third step: Neural Network, Support Vector Machine, K-nearest Neighbour and Gradient Boosting Decision Tree. The four individual methods belong to different types; we evaluate their performance by comparing the classification accuracy based on variable subsets selected in first two steps. Under the same threshold, we identified the most efficient method applied to the extensive banking data to predict the potential failure. These methods are widely used in predicting the distress in the financial market, such as the financial recess and the stock prices. They are suitable for our case in two ways, first they have no assumption on the statistical distribution of the data. Another advantage is their reliance on non-linear approaches, which can be more accurate when we have more complex data patterns. The details are introduced in the following content.

Neural Network (NN)

We applied a deep learning method, NN, that mimics human brain activities for classification. Given the selected variable set, and corresponding outputs (1 for failed banks and 0 for non-failed banks), it learns the transformation in the hidden layers concerning different weights on the predictors. Specifically, the assigned weights for variables are initially initialized and determine the importance, with larger ones contributing more to the outputs.

Then, the set of labeled training inputs is sent to the neural network to propagate through the hidden layers with the corresponding outputs, based on the weights fine-tuned in each hidden layer. The trained NN model is applied to the forecasting data and makes predictions.

It is noted by some studies that NN serves as more accurate and efficient classifiers than traditional methods such as logistic regression and linear discriminant analysis for prediction studies, especially with the extensive sample data (Kwon and Lee, 2015; Abdou et al., 2019). However, NN is criticized for its 'black-box' nature, indicating that generating the outputs through layers of neurons is non-transparent. Thus, errors in results are difficult to explain. At the same time, NN also requires a large number of input. Otherwise, the overfitting issues will lead to biased results.

Support Vector Machine (SVM)

In terms of dealing with the overfitting problem, we applied SVM (Vapnik, 1995), which performs better than NN in respect of requiring much fewer input data. SVM classifies the sample data by identifying a hyperplane that maximizes the margin between sample groups. Moreover, SVM addresses the non-linear classification problems by projecting the sample onto another higher dimensional space by using different kinds of kernel functions, including polynomial, Gaussian radial basis function, sigmoid, and hyperbolic tangent (Vapnik, 1998).

SVM is also one of the fastest robust algorithms with several appealing characteristics. It is flexible in the threshold separating the solvent banks from the insolvent banks, linear or non-linear. Compared with the NN with multiple solutions associated with the local optimal, SVM delivers a remarkable result by addressing the convex global optimization problem. However, the performance of SVM is sensitive to the outliers in the data. As the SVM decides the hyperplane based on the support vectors, data overlap of two target classes will lead to an inaccurate hyperplane that deviates from the correct position.

K-nearest Neighbor (KNN)

Meanwhile, we also use a non-parametric supervised technique KNN (Cover and Hart, 1967), to assign banks to failed or non-failed groups with their K nearest neighbors based on a distance measure such as Euclidean Minkowski Distance and Manhattan Distance. K is the number of members in each category. We utilize Euclidean distance measure and set K to be 10 in default. The advantage of KNN is its simplicity because it neither makes any assumptions on data nor uses the training data points to generalize the model framework. However, KNN is sensitive to the data quality, such as the existence of errors and outliers.

Gradient Boosting Decision Tree (GBDT)

To avoid outliers and redundant information issues, we applied GBDT from the ensemble tree methods. The basic idea of GBDT is to utilize a series of weak classifiers to a strong one. It trains each tree sequentially and reaches the classification target by successively minimizing the error. The loss function of GBDT is commonly regarding Negative binomial likelihood that is difficult to be optimized by the common gradient method. Thus, instead of weighting positive and negative samples, GBDT makes global convergence by using the negative gradient of the loss function to get the approximation of the loss in each training. It uses the least-squares function minimization to replace the challenging function minimizing issue, followed by a single parameter optimization based on the original criterion.

Suppose there are x variables with corresponding response function $F(x)$ that is the additive expansion based on the individual decision tree $h(x; a_m)$. Formally, function is written as $F(x) = \sum_{m=1}^{M} f_m(x) = \sum_{m=1}^{M} \beta_m h(x; a_m)$, where the β_m is estimated by minimizing the loss function $L(y, F(x)) = (y - F(x))^2$. Therefore, the GBDT algorithm is summarized as follow:

i. Initialize the response function to be a constant, $F_0(x) = \operatorname{argmin}_\beta \sum_{i=1}^{N} L(y_i, \beta)$.

ii. Compute the negative gradient: $m = 1, 2 \ldots M, i = 1, 2 \ldots N$

$$\tilde{y}_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{9}$$

iii. Fit the individual decision tree $h(x; a_m)$ to the target $\tilde{y}_{im}$

iv. Estimate the β_m with gradient descent and update the model:

$$\begin{gathered}\beta_m = \arg\min_\beta \textstyle\sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i; a_m)) \\ F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)\end{gathered} \tag{10}$$

v. The final output is $F(x) = F_M(x)$ GBDT is capable of handling different types of independent variables, capturing the multicollinearities, and fitting the non-linear relationship (Ke et al., 2017; Rao et al., 2019). It is noted that GBDT, as the second generation ensemble method, is superior to simple bagging methods such as Random Forests. The outcome of bagging methods depends on the majority vote, while the GBDT with proper hyperparameter tuning outperforms bagging by minimizing the error term to reach the optimal results.

5 Results

5.1 The Classification of Two Groups

We developed and applied a multi-stage model to see whether the applied indicators can accurately anticipate the bank failure and whether the model can efficiently forecast and distinguish between solvent and insolvent banks. Before proceeding to the forecasting, it is essential to conduct the variable selection in the first stage shown in figure A1 to reduce the influence of less irrelevant variables. In the first step, the mutual information selection generates a preliminary variable subset. The new subset consists of 136 (out of 405) predictors that are highly correlated with the dependent variable. In the second step, the Logistic Lasso regression applied to 136 predictors further shrinks the size of the subset to 42. In the procedure, the optimal lambda that is selected by cross-validation leads to the minimum estimation error. In the meantime, the RFs also select relevant features based on the MI subset. By comparing and utilizing results from two embedded methods, four variables having high feature importance that are identified by RFs are complemented to the variable subset. Ultimately, 54 indicators compose the comprehensive predictor subset with good explanation power.

Table 2 presents the categorical results of feature selection. Our findings suggest that variables related to capital, assets, and liquidity have a greater impact on explaining bank failure. The assets and capital have a negative relationship with bank insolvency as they are crucial factors for creating liquidity buffers for the bank to withstand shocks and crises, while uncollected loans are positively associated with bank risk due to default. Specifically, it is clear that capital directly related to risk resistance, such as risk-based capital and equity, are essential buffers during a crisis. In terms of assets and liquidity, a bank's loans have a complex association with its solvency. Well-performing loans can increase a bank's interest income, liquidity, and capital adequacy, which enhances the bank's solvency, while the provision of losses from loans is an expense for potentially uncollected loans and payments. Additionally, management and earnings, which are included to assess the health of banks, demonstrate the bank's financial performance and profitability. Effective management can facilitate sound financial decision-making, cost control, and profitability at a bank, which can contribute to its financial health. In terms of earnings, profitability is often viewed as an indicator of a bank's financial health, as it demonstrates the ability to generate sufficient revenue to cover expenses and remain solvent. In contrast, an unprofitable bank may struggle to maintain solvency and may potentially experience financial distress.

Table 2: Categories of Selected Variables In Second Step

Capital	**Asset**	**Management**	**Earnings**	**Liquidity**	**Sensitivity**
Capital Adequacy Ratio	Equity Growth Less Asset Growth	Net Operating Income to Average Asset	Return on Asset	Net Loans to Total Asset	Fair value of available-for-sale securities
Tier 1 Risk-Based Capital Ratio to Risk-weighted Assets	Noncurrent Loans and Leases to Gross Loans and Leases	Efficiency Ratio	Retain Earns to Average Total Equity	Net Loan to Core Deposits	
Tier 1 Leverage Ratio	Leases to Gross Loans and Leases			Total Deposits	
Equity to Assets Ratio	Average Total Loans and Leases			Loan to Deposit Shortfall	
Return on Equity	Loan and Lease Allowance				
Net Income					

In the third step, we applied the NN, SVM, KNN and GBDT to the in-sample data and out-of-sample data showed in time-window I. We also used the data from 2013-2019 for robustness check to figure out the consistency of the explanatory variables over time. We under-sample the training data to narrow the imbalance between the majority group and the minority group, while the test groups remain unchanged to avoid generating noises and overfitting as the result of manual adjustment. To evaluate the performance of each method based on different sample groups, we used the precision and sensitivity as basic metrics that are widely accepted for assessing measurement (Bekkar, Djemaa and Alitouche, 2013; Le and Viviani, 2018; Carmona, Climent and Momparler, 2019; Pinter et al., 2020; Chicco and

Jurman, 2020). The specificity and sensitivity are calculated by referring to the confusion matrix that summarizes the number of correct and incorrect predictions of classification problems. In particular, the confusion matrix shows the number of true positive (TP), true negative (TN) , false positive (FP) and false negative (FN), where TP refers to the number of positive cases, such as healthy banks, correctly identified as positive; TN is the number of negative cases such as failed bank correctly identified as negative; FP indicate the number of negative cases that are misidentified as the positive cases; FN shows the number of positive cases that are incorrectly identified as negative cases.

Formally, sensitivity is the proportion of true positive (TP) predictions to the aggregation of true positive (TP) and false negative (FN), and the specificity is the proportion of true negative (TN) to the sum of true negative (TN) and false positive (FP). Table 3 presents the confusion matrix composed of TP, FP, TN, FN.

Table 3: Confusion Matrix

True Class Predicted Class	Positive	Negative
Positive	TP	FP
Negative	FN	TN
Indicator	Sensitivity=$\frac{TP}{TP+FN}$	Specificity=$\frac{TN}{TN+FP}$

We adopted several models assessing measures that are calculated based on the sensitivity and specificity and assessed the performance of each method from two perspectives: the prediction accuracy and probability of misclassification. In particular, we are focusing on the following measures:

- Prediction Accuracy

i.Geometric Means (G-means): It evaluates the balance between class performances on the majority and minority groups:

$$G = \sqrt{\text{sensitivity} \ * \ \text{specificity}}$$

The score indicates the prediction performance of the positive cases. The poor performance of correctly recognizing the positive cases will lead to a low G-means value despite the high accuracy in correctly identifying the negative cases.

ii.Negative Likelihood Ratio (NLR): It is the ratio between the probability of predicting a case as false negative (FN) and true negative (TN).

$$NLR = \frac{1 - \text{ sensitivity}}{\text{specificity}}$$

The better performance will lead to lower NLR, which is an important indicator of the correctly identified insolvent cases in bank failure prediction.

iii.F1 Score: It evaluate the classification accuracy by combining the sensitivity and specificity of a classifier into a single metric by taking their harmonic mean. It is also used for comparing the performance of two classifiers when using the sensitivity and specificity are difficult to distinguish the better one.

$$F1 - \text{ score } = \frac{2 * \text{ sensitivity } * \text{ specificity}}{\text{sensitivity } + \text{ specificity}}$$

The higher F1 score is the result of better general performance of truly identifying the positive and negative cases, and the classifier outweighs its counterparts in classification based on the given variable set.

iv.Balanced Accuracy (BA): It is the simple average of sensitivity and specificity, which does not disregard the accuracy of the model in minority class.

$$BA = \frac{1}{2(\text{ sensitivity } + \text{ specificity })}$$

The conventional accuracy can be overoptimistic inflated as it takes advantage of the high prediction accuracy of the majority group. BA eliminates the adverse effect of the conventional accuracy by equally taking account of the majority group and the minority group.

- Probability of Misclassification

i.Youden's index (J): It is a linear combination of sensitivity and specificity.

$$J = \text{ sensitivity } - (1 - \text{ specificity })$$

The higher value of the J indicates the better ability of the classifier to avoid misclassification of banks.

ii.Area Under Curve (AUC): It measures the degree of the separability and the capability of model distinguishing between classes. It is graphically interpreted as the area below the receiver operating characteristic (ROC) curve. The common calculation method is the trapezoid that is based on the linear interpolation between each point on the ROC curve. The mathematical interpretation of AUC is the probability of ranking a random positive example more highly than a random negative example. In general, the AUC values varies between 0.5 and 1 and positively related to the classification performance. The AUC value above 0.8 indicates a very good performance with less misclassification.

iii.Matthew Correlation Coefficient (MCC): It is a contingency matrix method of calculating the Pearson product-moment correlation coefficient between actual and predicted values.

$$\text{MCC} = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The MCC is unaffected by the unbalanced data issue, and the value turns high if the binary classifier is able to correctly identify the majority of positive and negative cases.

Table 4 shows the results based on the in-sample data. In terms of accuracy, KNN and GBDT are superior to NN and SVM. It should be highlighted that KNN is the non-parametric machine learning method that does not have many assumptions about the underlying function, so it is more powerful and flexible. It is not surprising that GBDT delivers statistically significant forecasting power for bank failure, which is acknowledged by a strand of studies on the comparison of machine learning methods. When focusing on NLR, it shows that NN performs better than the KNN in identifying the insolvent banks, which is the main point of this study. However, it is possible that the performance of NN is optimistically inflated, which is the common pitfall of NN(Petropoulos et al., 2020).

Table 4: The Prediction Evaluation based on the In-sample Data

Method	Sensitivity	Specificity	F1-score	G-mean	NLR	BA	J	AUC	MCC
NN	0.850	0.634	0.726	0.734	0.236	0.742	0.484	0.766	0.211
SVM	0.895	0.621	0.733	0.745	0.169	0.758	0.515	0.797	0.244
KNN	0.752	0.801	0.776	0.776	0.310	0.776	0.553	0.772	0.263
GBDT	0.821	0.912	0.864	0.865	0.197	0.866	0.733	0.856	0.424

We examined the trained model based on the out-of-sample data (table 5) and find that GBDT aligns the best across the four methods. While, indicated by accuracy indicators including F1-score and G-means, KNN ranking the second is found performing superior to the SVM. While in terms of avoiding the misclassification, the NLR, AUC and J-index illustrate that KNN is underperformed than SVM, leading to a higher probability of classifying the true cases to the wrong group. Regarding the out-of-time performance, the KNN and GBDT align the best fit, with the former exhibiting marginally better performance in 5 criteria and the latter delivering marginally overperformance in all criteria. Additionally, in the out-of-time performance evaluation, NN has the poorest performance in all the criteria.

Table 5: The Prediction Evaluation based on the Out-of-sample Data

Methods	Sensitivity	Specificity	F1-score	G-mean	NLR	BA	J	AUC	MCC
NN	0.878	0.647	0.745	0.754	0.189	0.762	0.525	0.774	0.239
SVM	0.932	0.662	0.774	0.785	0.103	0.797	0.594	0.795	0.266
KNN	0.809	0.783	0.796	0.796	0.244	0.796	0.591	0.770	0.284
GBDT	0.813	0.911	0.859	0.860	0.206	0.862	0.723	0.864	0.429

Table 6: The Prediction Evaluation based on the 13-19 Data

Methods	Sensitivity	Specificity	F1-score	G-mean	NLR	BA	J	AUC	MCC
NN	0.665	0.793	0.723	0.726	0.423	0.729	0.458	0.729	0.435
SVM	0.782	0.838	0.809	0.809	0.260	0.810	0.620	0.810	0.589
KNN	0.730	0.901	0.806	0.811	0.300	0.815	0.631	0.815	0.633
GBDT	0.751	0.977	0.849	0.856	0.255	0.864	0.728	0.864	0.784

In general, it is evident that GBDT exhibits has better prediction power than the rest models, and the performance provided is stable and consistent across all sample groups. It is an adequate tool obtained for future bank risk estimation. In contrast, NN performs poorly both in the "in-sample" and "out-of-time" samples. Regarding the comparisons of three samples, all models perform relatively well in the first time window, while the GBDT and KNN provide pervasively good results across different time windows. The consistent performance of the trained models across time also denotes that the selected variables are essential in signaling the bank risk-taking behavior in both crisis and stable period.

It is acknowledged that the unanticipated bank failure can trigger large systemic risk in the financial market. Due to contagion effects, the cost of bank failure is much higher than the default of other financial institutions. The increasing demanding for money from depositors leads to serious liquidity issues and the declining money supply to bank financing corporates results in a slowing economic development. On the other hand, the bail-out operation from the government transfers the overloading costs to taxpayers.Thus, it is imperative for supervisors to capture early signals of the insolvency based on the accurate prediction results, and then assign the pre-cautions to resist the destabilization of the economy in time.

5.2 The Classification of Multi-Groups

In the process, we also answer some additional questions based on the risk level of bank failure that is barely mentioned in the existing literature. Most of the studies have focused on the technical predictive procession of two classes, the failed and the non-failed, while the notion of how much risk can banks bear without default is still unclear from forecasting. To get a clearer picture, we use the classical z score(Roy, 1952; Laeven and Levine, 2009) to evaluate the bank default risks. Table 7 shows the results of the default risk, summarized by groups. The mean and quantile results show the distinct default risks between the failed and non-failed banks. However, whether the risk level difference is significant and whether the failed banks always have higher risk levels than the non-failed banks are still unknown.

Table 7: Description of Default Risk by Groups

Variables	Obs	Mean	Std. Dev.	p25	p50	p75	Skew.	Kurt.
ID = 1	6560	21.907	18.54	6.116	20.182	31.937	1.471	8.002
ID = 0	167775	37.305	25.229	22.911	32.313	45.113	5.218	105.189

To have a clearer perception, we reclassify banks based on bank solvency and level of risk to default to investigate the boundary between high risk and failure. We use the p50 z-score of failed banks as the threshold to divide the two-class banks into four classes. The banks with z-scores less than the value of p50 are labeled with higher risk, and the remaining are labeled with lower risk. If the risk level difference is significant, banks can be accurately classified into different risk level groups given the indicators of bank performance. Thus, the influence of bank risk level on bank failure will be evaluated by classification accuracy.

In terms of classification method, we use GBDT which has the best performance in the previous section, and the same 54 selected features. The first row of table 8 shows the classification evaluation results. The F1-score, G-mean, BA, and J score indicate that GBDT remains the good performance of classification based on the new labels and selected predictors. The prediction accuracy proves that the risk level difference is significant within and between the failed and non-failed banks, and the non-failed banks can have more risk than the failed banks. With adequate liquidity and capital, the non-failed banks are capable of bearing more risks, which helps them to survive during the crisis.

Table 8: Prediction Evaluation of Multi-groups during the Financial Crisis

Assessment	Sensitivity	Specificity	F1-Score	G-mean	NLR	BA	J
Four groups	0.782	0.882	0.829	0.831	0.247	0.832	0.664
Six groups	0.719	0.863	0.784	0.788	0.326	0.791	0.582

Furthermore, we also reclassify the banks more specifically into six groups based on the p50 z-score of both failed banks and non-failed banks. The new labels describe banks that failed or non-failed with lower, medium, or higher default risks. The second row in table 8 shows the classification results. The decline of evaluation scores compared with the first row can be attributed to either insignificant differences between medium and higher risk levels or insufficient data in each group for model training due to an increased number of subgroups. We also did the robust test based on the out-of-sample data which covers 2013-2019 (see table A3). The results are consistent with evaluations of the crisis period.

6 Conclusions

Bank failure prediction is an important topic due to the intermediation role of banks. The study intends to increase forecasting efficiency to provide early warnings of bankruptcy. Our analysis used a sample of U.S. data consisting of 6801 banks from 2006-2019. Based on CAMELS, we applied mutual information selection for examining the correlation, and embedded feature selection methods, including Logistic Lasso and Random Forest feature importance, for further variable selection. Then, we examine the performance of NN, SVM, KNN, and GBDT in classification, employing the selected variable subset.

The variable selection results show that the combination of the correlation test and the embedded methods decreased the number of variables to 54. Moreover, we found that indicators related to capital, asset and liquidity are essential and consistent in explaining bank failures over time and can be devoted to long-term forecasting.

The comparative prediction results show that the best fit model is GBDT which delivers the most accurate results. Besides, KNN has relatively more prediction power when evaluating samples of crisis period and NN has the worst performance. In terms of practical value, GBDT is an outstanding classification tool to help supervisors achieve the optimal possible accuracy when setting an early warning system for bank insolvency. We contribute to the literature by examining different types of machine learning models based on long-term forecasting and validating the consistency of the selected variables for further bank failure prediction. Additionally, we capture the influence of the default risk level on the bank failure and identify the significant difference in level risks within and between failed and non-failed banks. The results indicate that the combined effect of adequate liquidity and capital is the key to highly risky banks overcoming the shocks such as bank-run, which is meaningful for regulation and supervision.

In respect of the limitations, CAMELS may not fully describe the risks due to the nature of banks' operation and practice strategies. Further research will explore more features such as market data, corporate governance, diversification, non-financial activities, green finance regulations, macroeconomic factors, and customer-related indicators. Further study will focus on adding diversification indicators and completing hybrid prediction models.

References

Abdou, H.A., Mitra, S., Fry, J. and Elamer, A.A., 2019. Would two-stage scoring models alleviate bank exposure to bad debt? *Expert systems with applications*, 128, pp.1–13.

Apergis, N. and Payne, J.E., 2013. European banking authority stress tests and bank failure: Evidence from credit risk and macroeconomic factors. *Banking and finance review*, 5, p.2.

Audrino, F., Kostrov, A. and Ortega, J.P., 2019. Predicting us bank failures with midas logit models. *Journal of financial and quantitative analysis*, 54(6), pp.2575–2603.

Bank, T.W. and Fund, T.I.M., 2005. *Financial sector assessment - a handbook.* Washington: DC.

Barth, J.R., Trimbath, S. and Yago, G.E., 2006. *The savings and loan crisis: Lessons from a regulatory failure (vol. 5).* Science and Business Media: Springer.

Bekkar, M., Djemaa, H.K. and Alitouche, T.A., 2013. Evaluation measures for models assessment over imbalanced data sets. *J inf eng appl*, 3, p.10.

Beltratti, A. and Stulz, R.M., 2012. The credit crisis around the globe: Why did some banks perform better? *Journal of financial economics*, 105(1), pp.1–17.

Bennasar, M., Hicks, Y. and Setchi, R., 2015. Feature selection using joint mutual information maximisation. *Expert systems with applications*, 42(22), pp.8520–8532.

Berger, A. N., .B.C.H., 2013. How does capital affect bank performance during financial crises? *Journal of financial economics*, 109(1), pp.146–176.

Berger, A.N., Imbierowicz, B. and Rauch, C., 2016. The roles of corporate governance in bank failures during the recent financial crisis. *Journal of money, credit and banking*, 48(4), pp.729–770.

Bertay, A.C., Demirg"uç-Kunt, A. and Huizinga, H., 2013. Do we need big banks? *Evidence on performance, strategy, and market discipline*, 22(4), pp.532–558.

Boyacioglu, M.A. and Kara, Y.a., 2009. and baykan, "O. *K. predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (sdif) transferred banks in turkey*, 36(2), pp.3355–3366.

Boyallian, P. and Ruiz-Verdú, P., 2018. Leverage, ceo risk-taking incentives, and bank failure during the 2007–10 financial crisis. *Review of finance*, 22(5), pp.1763–1805.

Canbas, S., Cabuk, A. and Kilic, S.B., 2005. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The turkish case[j]. *European journal of operational research*, 166(2), pp.528–546.

Carmona, P., Climent, F. and Momparler, A., 2019. Predicting failure in the us banking sector: An extreme gradient boosting approach. *International review of economics and finance*, 61, pp.304–323.

Chernykh, L. and Kotomin, V., 2022. Risk-based deposit insurance, deposit rates and bank failures: Evidence from russia. *Journal of banking and finance*, 138(10648), p.3.

Chicco, D. and Jurman, G., 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *Bmc genomics*, 21(1), pp.1–13.

Climent, F., Momparler, A. and Carmona, P., 2019. Anticipating bank distress in the eurozone: An extreme gradient boosting approach. *Journal of business research*, 101, pp.885–896.

Cole, R.A. and White, L.J., 2012. Déjà vu all over again: The causes of us commercial bank failures this time around. *Journal of financial services research*, 42(1-2), pp.5–29.

Cole, R.A. and Wu, Q., 2014. Hazard versus probit in predicting us bank failures: a regulatory perspective over two crises. *Available at ht tps://ssrn. com/abstract*, 1460526.

Curry, T. and Shibut, L., 2000. The cost of the savings and loan crisis: Truth and consequences. *Fdic banking rev.*, 13, p.26.

Demyanyk, Y. and Hasan, I., 2010. Financial crises and bank failures: A review of prediction methods. *Omega*, 38(5), pp.315–324.

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E. and Zhang, W., 2022. *Operational research and artificial intelligence methods in banking.* European Journal of Operational Research.

Ecer, F., 2013. Comparing the bank failure prediction performance of neural networks and support vector machines: The turkish case. *Economic research-ekonomska istraživanja*, 26(3), pp.81–98.

Ekinci, A. and Erdal, H.İ., 2017. Forecasting bank failure: Base learners, ensembles and hybrid ensembles. *Computational economics*, 49(4), pp.677–686.

Erdal, H.I. and Ekinci, A., 2013. A comparison of various artificial intelligence methods in the prediction of bank failures. *Computational economics*, 42(2), pp.199–215.

Fahlenbrach, R., Prilmeier, R. and Stulz, R.M., 2012. This time is the same: Using bank performance in 1998 to explain bank performance during the recent financial crisis. *The journal of finance*, 67(6), pp.2139–2185.

Fethi, M.D. and Pasiouras, F., 2010. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. *European journal of operational research*, 204(2), pp.189–198.

Ghaddar, B. and Naoum-Sawaya, J., 2018. High dimensional data classification and feature selection using support vector machines. *European journal of operational research*, 265(3), pp.993–1004.

Goenner, C.F., 2020. Uncertain times and early predictions of bank failure. *Financial review*, 55(4), pp.583–601.

Gogas, P., Papadimitriou, T. and Agrapetidou, A., 2018. Forecasting bank failures and stress testing: A machine learning approach. *International journal of forecasting*, 34(3), pp.440–455.

Jin, J.Y., Kanagaretnam, K. and Lobo, G.J., 2011. Ability of accounting and audit quality variables to predict bank failure during the financial crisis. *Journal of banking and finance*, 35(11), pp.2811–2819.

Jin, J.Y., Kanagaretnam, K., Lobo, G.J. and Mathieu, R., 2013. Impact of fdicia internal controls on bank risk taking. *Journal of banking and finance*, 37(2), pp.614–624.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W. and Ma, W., 2017. ... and liu, t. *Y. lightgbm: A highly efficient gradient boosting decision tree*, 30.

K"ohler, M., 2015. Which banks are more risky? the impact of business models on bank stability. *Journal of financial stability*, 16, pp.195–212.

Kwon, H.B. and Lee, J., 2015. Two-stage production modeling of large us banks: A dea-neural network approach. *Expert systems with applications*, 42(19), pp.6758–6766.

Laeven, L. and Levine, R., 2009. Bank governance, regulation and risk taking. *Journal of financial economics*, 93(2), pp.259–275.

Le, H.H. and Viviani, J.L., 2018. Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. *Research in international business and finance*, 44, pp.16–25.

Lopez, J.A., 1999. *Using camels ratings to monitor bank conditions.* FRBSF Economic Letter.

Maldonado, S., Pérez, J. and Bravo, C., 2017. Cost-based feature selection for support vector machines: An application in credit scoring. *European journal of operational research*, 261(2), pp.656–665.

Manthoulis, G., Doumpos, M., Zopounidis, C. and Galariotis, E., 2020. An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for us banks. *European journal of operational research*, 282(2), pp.786–801.

Manthoulis, G., Doumpos, M., Zopounidis, C., Galariotis, E. and Baourakis, G., 2021. Bank failure prediction: A comparison of machine learning approaches. *Financial risk management and modeling . , cham.* pp.349–366.

Mare, D.S., 2015. Contribution of macroeconomic factors to the prediction of small bank failures. *Journal of international financial markets, institutions and money*, 39, pp.25–39.

Markovitch, S. and Rosenstein, D., 2002. Feature generation using general constructor functions. *Machine learning*, 49(1), pp.59–98.

Mayes, D.G. and Stremmel, H., 2012. *The effectiveness of capital adequacy measures in predicting bank distress.* In 2013 Financial Markets and Corporate Governance Conference.

Meier, L., Van De Geer, S. and B"uhlmann, P., 2008. The group lasso for logistic regression. *Journal of the royal statistical society: Series b (statistical methodology)*, 70(1), pp.53–71.

Mili, M., Khayati, A. and Khouaja, A., 2019. *Do bank independency and diversification affect bank failures in europe?* Review of Accounting and Finance.

Ng, G.S., Quek, C. and Jiang, H., 2008. Fcmac-ews: A bank failure early warning system based on a novel localized pattern learning and semantically associative fuzzy neural network. *Expert systems with applications*, 34(2), pp.989–1003.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Vlachogiannakis, N.E., 2020. Predicting bank insolvencies using machine learning techniques. *International journal of forecasting*, 36(3), pp.1092–1113.

Pham, T.T.X. and Ho, H.T., 2021. *Using boosting algorithms to predict bank failure: An untold story.* International Review of Economics and Finance.

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. and Gloaguen, R., 2020. Covid-19 pandemic prediction for hungary; a hybrid machine learning approach. *Mathematics*, 8(6), p.890.

Ramirez, C.D. and Shively, P.A., 2012. The effect of bank failures on economic activity: Evidence from us states in the early 20th century. *Journal of money, credit and banking*, 44(2-3), pp.433–455.

Rao, H., Shi, X., Rodrigue, A.K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X. and Gu, L., 2019. Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied soft computing*, 74, pp.634–642.

Rapach, D.E., Strauss, J.K. and Zhou, G., 2013. International stock return predictability: What is the role of the united states? *The journal of finance*, 68(4), pp.1633–1662.

Roy, A.D., 1952. Safety first and the holding of assets. *Econometrica: Journal of the econometric society*, pp.431–449.

Sarkar, S. and Sriram, R.S., 2001. Bayesian models for early warning of bank failures. *Management science*, 47(11), pp.1457–1475.

Serrano-Cinca, C., Fuertes-Callén, Y., Gutiérrez-Nieto, B. and Cuellar-Fernández, B., 2014. Path modelling to bankruptcy: causes and symptoms of the banking crisis. *Applied economics*, 46(31), pp.3798–3811.

Shaban, M. and James, G.A., 2018. The effects of ownership change on bank performance and risk exposure: Evidence from indonesia. *Journal of banking and finance*, 88, pp.483–497.

Shannon, C.E., 1948. A mathematical theory of communication. *The bell system technical journal*, 27(3), pp.379–423.

Sun, J., Wu, D. and Zhao, X., 2018. Systematic risk factors and bank failures. *Journal of economics and business*, 98, pp.1–18.

Tam, K.Y. and Kiang, M.Y., 1992. Managerial applications of neural networks: the case of bank failure predictions. *Management science*, 38(7), pp.926–947.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the royal statistical society: Series b (methodological)*, 58(1), pp.267–288.

Unler, A. and Murat, A., 2010. A discrete particle swarm optimization method for feature selection in binary classification problems. *European journal of operational research*, 206(3), pp.528–539.

Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K., 2020. Intelligent hybrid model for financial crisis prediction using machine learning techniques. *Information systems and e-business management*, 18(4), pp.617–645.

Vapnik, V., 1995. *The nature of statistical learning theory.* New York: Springer Verlag.

Vapnik, V., 1998. *Statistical learning theory.* NY: Wiley.

Vazquez, F. and Federico, P., 2015. Bank funding structures and risk: Evidence from the global financial crisis. *Journal of banking and finance*, 61, pp.1–14.

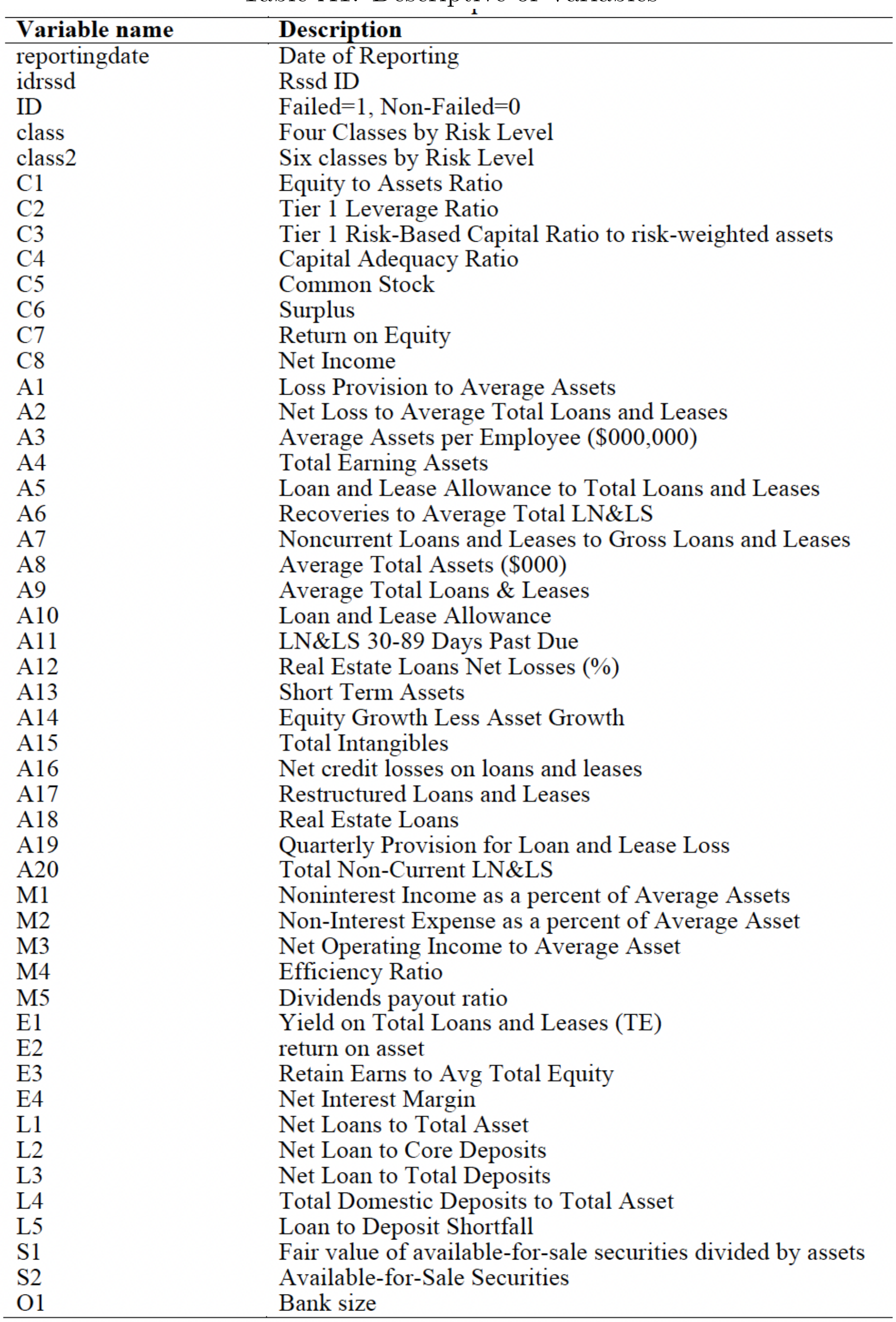
Wulandari, Y. and Kusairi, S., 2017. The impact of macroeconomic and internal factors on banking distress. *International journal of economics and financial issues*, 7(3), pp.429–436.

Zhao, H., Sinha, A.P. and Ge, W., 2009. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert systems with applications*, 36(2), pp.2633–2644.

Zou, H., 2006. The adaptive lasso and its oracle properties. *Journal of the american statistical association*, 101(476), pp.1418–1429.

A Appendix

Table A1: Descriptive of Variables



Variable name	Description
reportingdate	Date of Reporting
idrssd	Rssd ID
ID	Failed=1, Non-Failed=0
class	Four Classes by Risk Level
class2	Six classes by Risk Level
C1	Equity to Assets Ratio
C2	Tier 1 Leverage Ratio
C3	Tier 1 Risk-Based Capital Ratio to risk-weighted assets
C4	Capital Adequacy Ratio
C5	Common Stock
C6	Surplus
C7	Return on Equity
C8	Net Income
A1	Loss Provision to Average Assets
A2	Net Loss to Average Total Loans and Leases
A3	Average Assets per Employee ($000,000)
A4	Total Earning Assets
A5	Loan and Lease Allowance to Total Loans and Leases
A6	Recoveries to Average Total LN&LS
A7	Noncurrent Loans and Leases to Gross Loans and Leases
A8	Average Total Assets ($000)
A9	Average Total Loans & Leases
A10	Loan and Lease Allowance
A11	LN&LS 30-89 Days Past Due
A12	Real Estate Loans Net Losses (%)
A13	Short Term Assets
A14	Equity Growth Less Asset Growth
A15	Total Intangibles
A16	Net credit losses on loans and leases
A17	Restructured Loans and Leases
A18	Real Estate Loans
A19	Quarterly Provision for Loan and Lease Loss
A20	Total Non-Current LN&LS
M1	Noninterest Income as a percent of Average Assets
M2	Non-Interest Expense as a percent of Average Asset
M3	Net Operating Income to Average Asset
M4	Efficiency Ratio
M5	Dividends payout ratio
E1	Yield on Total Loans and Leases (TE)
E2	return on asset
E3	Retain Earns to Avg Total Equity
E4	Net Interest Margin
L1	Net Loans to Total Asset
L2	Net Loan to Core Deposits
L3	Net Loan to Total Deposits
L4	Total Domestic Deposits to Total Asset
L5	Loan to Deposit Shortfall
S1	Fair value of available-for-sale securities divided by assets
S2	Available-for-Sale Securities
O1	Bank size

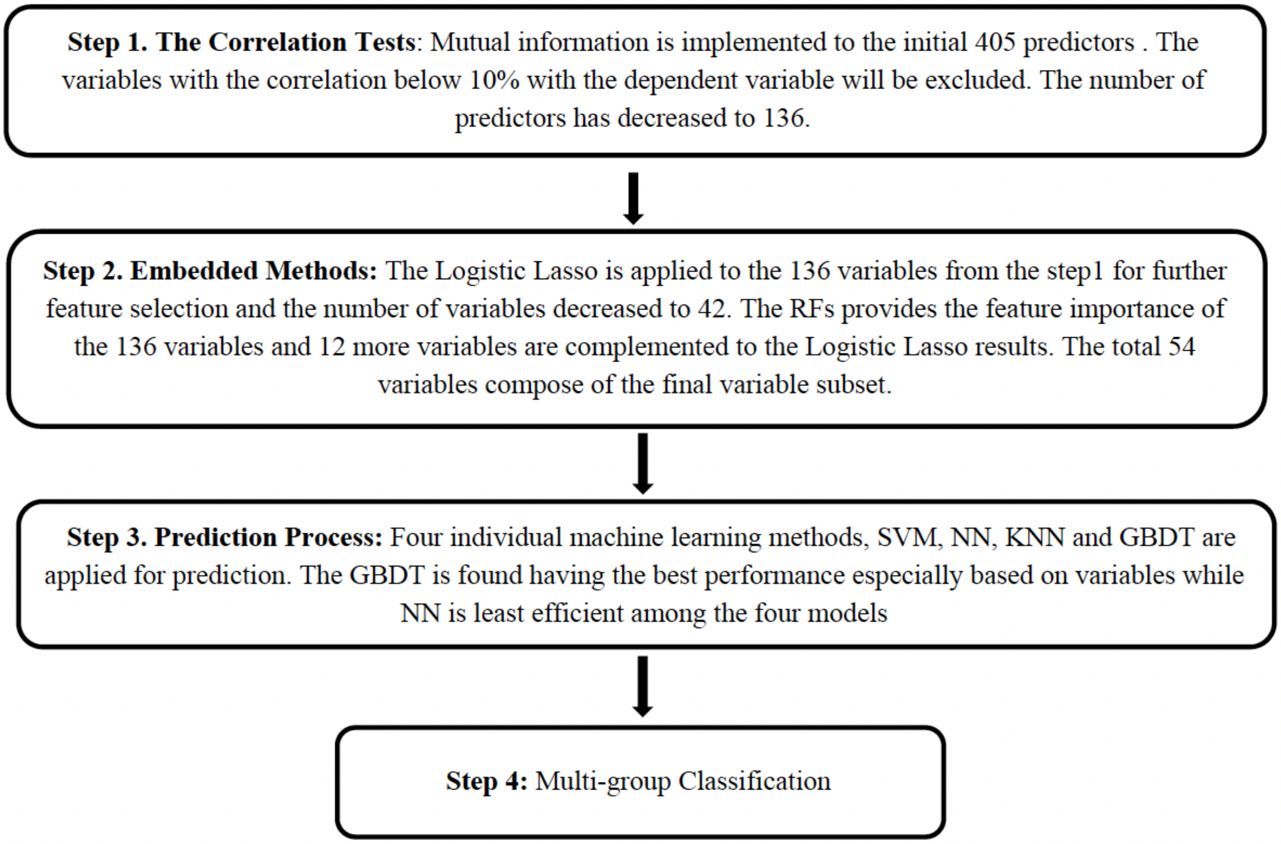
Table A2: Summary of Variables by Groups

	ID==1					ID==0				
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
C1	6,560	0.0803218	0.03428	-0.118	0.341669	167,775	0.108117	0.036	-0.0105	0.908878
C2	6,560	7.752822	3.366379	-11.41	46.74	167,775	10.32164	3.672	-4.79	518.26
C3	6,560	9.692756	4.289701	-16.53	73.1512	167,775	15.92875	11.675	-6.76	3349.17
C4	6,560	10.91879	4.289667	-16.53	74.5246	167,775	17.0774	12.029	-6.76	3564.17
C5	6,560	4246.175	14403.03	0	235815	167,775	4069.068	52488.2	0	3020043
C6	6,560	38788.89	123085.5	-202	2495494	167,775	116095.7	2470399	0	1.45E+08
C7	6,560	-27.76464	198.4061	-11095.8	104.5	167,775	8.010078	33.345	-754.43	11510.15
C8	6,560	-3272.058	31652.15	-1085667	280117	167,775	7415.888	188692	-6E+06	1.66E+07
A1	6,560	1.370777	2.451807	-4.68	76.08	167,775	0.357945	0.769	-14.74	72.5
A2	6,560	1.455134	2.803245	-3.38	76.34	167,775	0.439271	1.119	-28.32	117.76
A3	6,560	4.70939	2.606549	1.07	31.8	167,775	4.513195	35.4082	0.14	9004.27
A4	6,560	617654.4	1645819	9317	2.42E+07	167,775	1486989	3E+07	668	1.69E+09
A5	6,560	2.237223	1.730552	0.14	25.81	167,775	1.569021	0.928	0	24.19
A6	6,560	0.0946006	0.38394	0	18.8	167,775	0.098185	0.290	0	28.79
A7	6,560	5.937508	7.039161	0	71.17	167,775	1.884805	2.551	0	81.98
A8	6,560	672797.1	1803985	10762	2.76E+07	167,775	1651238	3.32E+07	3550	1.81E+09
A9	6,560	485346.9	1196910	6282	1.92E+07	167,775	919582.5	1.58E+07	238	7.58E+08
A10	6,560	10382.71	27207.36	34	495000	167,775	19711.8	437765	0	2.60E+07
A11	6,560	9986.245	26940.81	0	417059	167,775	12479.32	264713	0	1.79E+07
A12	6,560	1.35803	2.841163	-4.49	53.8	167,775	0.350397	1.1086	-52.14	95.33
A13	6,560	225451.9	535804.1	221	7831792	167,775	416520.1	9253374	0	5.69E+08
A14	6,560	-12.00813	33.38266	-206.34	826.95	167,775	1.569777	34.1126	-9141.8	2819.83
A15	6,560	1.219195	262.2089	-21130.8	1823.73	167,775	4.328702	17.4948	-5850	488.77
A16	6,560	4583.412	21551.13	-3287	642787	167,775	7014.838	188661	-13000	2.00E+07
A17	6,560	0.9983399	2.48641	0	31.83	167,775	0.418411	1.16877	0	30.72
A18	6,560	412403.1	1034949	1320	1.68E+07	167,775	542147	8540919	0	4.78E+08
A19	6,560	3297.022	15453.51	-64357	451883	167,775	3264.492	85094.7	-227000	9790785
A20	6,560	29554.66	111668.7	0	2625116	167,775	30715.66	861092	0	6.59E+07
M1	6,560	0.5738262	1.163007	-15.52	53.23	167,775	0.824883	3.11024	-23.01	987.52
M2	6,560	3.336273	1.822637	0.61	92.77	167,775	3.078276	1.46665	0	86.53
M3	6,560	-1.473114	4.915382	-72.53	34.09	167,775	0.760669	1.6224	-91.55	56.87
M4	6,560	105.6328	475.4702	-2295.73	25375	167,775	68.67084	44.4694	-13662	1987.5
M5	6,560	15.84696	471.7823	-28000	13828.57	167,775	41.33221	2837.68	-1E+06	50100
E1	6,560	7.217316	1.86291	2.32	91.1	167,775	6.894816	1.28511	0	63.11
E2	6,560	-0.825282	3.399322	-72.53	6.51	167,775	0.851074	2.47115	-71.36	701.01
E3	6,560	-37.67811	663.5181	-36030.7	22438.81	167,775	1.816456	22.3004	-3564	1123.59
E4	6,560	3.679726	1.267439	-0.74	41.65	167,775	3.91977	0.94353	-1.41	32.27
L1	6,560	0.7291743	0.11462	0.232778	0.97165	167,775	0.626332	0.15207	0.0007	0.993702
L2	6,560	1.380267	23.5531	-1613.9	645.1805	167,775	26.79062	2768.81	-1697.1	524963.7
L3	6,560	0.887647	0.28891	0.267	10.76491	167,775	1.997224	287.698	0.00101	116308
L4	6,560	0.8376407	0.088546	0.060	1.101302	167,775	0.8275	0.07898	0	0.987015
L5	6,560	-0.1089021	0.146896	-0.640	0.589423	167,775	-0.20227	0.17355	-0.9439	0.992894
S1	6,560	11.03815	9.399171	0	65.84	167,775	19.19632	14.404	0	95.06
S2	6,560	89229.55	335844.4	0	5183085	167,775	286435.4	5795243	0	3.70E+08
O1	6,560	12.47693	1.240982	9.305	17.12248	167,775	12.03647	1.31218	8.02715	21.36342

Table A3: Prediction Evaluation of Multi-groups based on the Out-of-sample Data

Assessment	Sensitivity	Specificity	F1-Score	G-mean	NLR	BA	J
Four groups	0.617	0.922	0.739	0.754	0.415	0.770	0.539
Six groups	0.497	0.938	0.650	0.683	0.536	0.717	0.435

Figure A1: The Flow Chart of Research Process



Step 1. The Correlation Tests: Mutual information is implemented to the initial 405 predictors . The variables with the correlation below 10% with the dependent variable will be excluded. The number of predictors has decreased to 136.

Step 2. Embedded Methods: The Logistic Lasso is applied to the 136 variables from the step1 for further feature selection and the number of variables decreased to 42. The RFs provides the feature importance of the 136 variables and 12 more variables are complemented to the Logistic Lasso results. The total 54 variables compose of the final variable subset.

Step 3. Prediction Process: Four individual machine learning methods, SVM, NN, KNN and GBDT are applied for prediction. The GBDT is found having the best performance especially based on variables while NN is least efficient among the four models

Step 4: Multi-group Classification