Data Science in Finance and Economics

DSFE, 1(2): 180–195.
DOI: 10.3934/DSFE.2021010
Received: 08 August 2021
Accepted: 27 September 2021
Published: 11 October 2021

http://www.aimspress.com/journal/dsfe

*Research article*

# Bankruptcy prediction using machine learning and an application to the case of the COVID-19 recession

**Aditya Narvekar[1,*] and Debashis Guha[2]**

[1] Department of Data Science, SP Jain School of Global Management, Sydney, Australia
[2] Department of Data Science, SP Jain School of Global Management, Mumbai, India

*** Correspondence:** Email: aditya.narvekar@gmail.com; Tel: +61467524385.

**Abstract:** Bankruptcy prediction is an important problem in finance, since successful predictions would allow stakeholders to take early actions to limit their economic losses. In recent years many studies have explored the application of machine learning models to bankruptcy prediction with financial ratios as predictors. This study extends this research by applying machine learning techniques to a quarterly data set covering financial ratios for a large sample of public U.S. firms from 1970–2019. We find that tree-based ensemble methods, especially XGBoost, can achieve a high degree of accuracy in out-of-sample bankruptcy prediction. We next apply our best model, using XGBoost, to the problem of predicting the overall bankruptcy rate in USA in the second half of 2020, after the COVID-19 pandemic had necessitated a lockdown, leading to a deep recession. Our model supports the prediction, made by leading economists, that the rate of bankruptcies will rise substantially in 2020, but it also suggests that this elevated level will not be much higher than 2010.

**Keywords:** bankruptcy prediction; machine learning; multivariate discriminant analysis; logit and probit models; big data; ensemble models; XGBoost; Random Forests; support vector machines

**JEL Codes:** O30, E37, E60

## 1. Introduction

Bankruptcy prediction is the problem of detecting financial distress in businesses which will lead to eventual bankruptcy. Bankruptcy prediction has been studied since at least 1930s. The early models of bankruptcy prediction employed univariate statistical models over financial ratios. The univariate models were followed by multi-variate statistical models such as the famous Altman Z-score model. The recent advances in the field of Machine learning have led to the adoption of Machine learning algorithms for bankruptcy prediction. Machine Learning methods are increasingly being used for bankruptcy prediction using financial ratios. A study by Barboza, Kimura and Altman found that Machine Learning models can outperform classical statistical models like multiple discriminant analysis (MDA) by a significant margin in bankruptcy prediction (Barboza et al., 2017).

### *1.1. Significance of bankruptcy prediction*

Bankruptcy prediction is an important for modern economies because early warnings of bankrupt help not only the investor but also public policy makers to take proactive steps to minimize the impact of bankruptcies. The reasons that add to the significance of bankruptcy prediction are as follows:

(1). Better allocation of resources

Institutional investors, banks, lenders, retail investors are always looking at information that predicts financial distress in publicly traded firms. Early prediction of bankruptcy helps not only the investors and lenders but also the managers of a firm to take corrective action thereby conserving scare economic resources. Efficient allocation of capital is the cornerstone of growth in modern economies.

(2). Input to policy makers

Accurate prediction of bankruptcies of businesses and individuals before they happen gives law makers and policy makers some additional time to alleviate systemic issues that might be causing the bankruptcies. Indeed, with bankruptcies taking center stage in political discourse of many countries, the accurate prediction of bankruptcy is a key input for politicians, bureaucrats and in general for anyone who is making public policy.

(3). Corrective action for business managers

The early prediction of bankruptcy is likely to highlight business issues thereby giving the company's manager additional time to make decisions that will help avoid bankruptcy. This effect is likely to be more profound in public companies where the management has a fiduciary duty to the shareholders.

(4). Identification of sector wide problems

Bankruptcy prediction models that flag firms belonging to a certain sector are likely to be a leading indicator of an upcoming downturn in a certain sector of an economy. With robust models, the business managers and government policy makers would become aware and take corrective action to limit the magnitude and intensity of the downturn in the specific sector. Industry groups in turn has been shown to significantly effect forecasting models (Chava and Jarrow, 2004).

(5). Signal to Investors

Investors can make better and more informed decisions based on the prediction of bankruptcy models. This not only forces the management of firms to take corrective action but also helps to soften the overall

economic fallout that results from the bankruptcies. Empirical studies have shown that investment opportunities are significantly related to likelihood of bankruptcy (Lyandres & Zhdanov, 2007).

(6). Relation to adjacent problems

Bankruptcy prediction is often the first step used by ratings agencies to detect financial distress in firms. Based on the predictions of bankruptcy models, ratings agencies investigate and assess credit risk. Getting flagged by bankruptcy prediction models is often the first step that triggers the process of revising credit ratings. A literature survey covering 2000–2013 demonstrates the close relation between bankruptcy prediction and credit risk (García et al., 2015).

### *1.2. Comparison with past work*

Most past studies in bankruptcy prediction including those using Machine Learning have used a relatively small sample of firms and a small number of financial ratios. This study distinguishes itself by using a much larger dataset having data for 21,114 U.S. firms (samples) and 57 financial ratios (features). Our dataset covers US firms from 1970 to 2020. Bankruptcy prediction models have been researched and built since the 1960s. The models built from 1960 to 1990 were primarily statistical models such as univariate, multiple discriminant analysis and logit and probit models. Starting from 1990s machine learning models started outperforming statistical models. Since this study applies the most popular contemporary machine learning algorithms using a big data set, we will compare our model with the machine learning models built since the 1990s. A full listing outlining the comparison with past machine learning studies and models for bankruptcy prediction is shown in the Table 1.

In this study we have used three popular machine learning techniques—Random Forest, Support Vector Machines, and XGBoost to construct forecasting models. We find that Machine Learning models perform very well, with XGBoost being the most successful technique that achieves an accuracy score of more than 99% in out of sample testing.

We also apply our XGBoost model to an important current issue, the task of predicting bankruptcies during the second half of 2020. The depth of the recession caused by the lockdowns that have been imposed to contain the COVID-19 pandemic has raised worries that corporate bankruptcies may rise substantially in the near future. According to a report in the New York Times (2020), Edward Altman, a pioneer of bankruptcy prediction research, and the creator of the famous Z score model, expects a “tsunami of bankruptcies” that will exceed the number of bankruptcies that followed the 2008 financial crisis. The result from our Machine Learning model confirms Prof Altman’s fears that corporate bankruptcies will rise substantially in late 2020 and equal the highs seen during the 2008-09 recession. However, this study finds that the elevated level of bankruptcies will not be significantly different from 2010.

**Table 1.** Past ML studies.

| Author | Machine learning models used | Number of features (ratios) | Size of training set |
|---|---|---|---|
| Wilson and Sharda (1994) | Shallow neural network, multi-discriminant analysis | 5 ratios (same ratios used by Altman) | 65—Bankrupt firms<br>64—non-bankrupt firms<br>Total—169 firms |
| Min and Lee (2005) | Support vector machine (SVM), Multi-discriminant analysis, Logistic Regression, Shallow Neural Network | 38 ratios used to compute 2 principal components. | 944—Bankrupt firms<br>944—non-bankrupt firms<br>Total—1888 firms |
| Fedorova, Gilenko and Dovzhenko (2013) | Adaptive Boost, Artificial Neural Network, Adaptive Boost combined with Neural Networks | 75 ratios | 444—Bankrupt firms<br>444—non-bankrupt firms<br>Total—888 firms |
| Shi, Xi, Ma, Hu (2009) | Bagging ensemble of Artificial Neural Networks (ANNs), Artificial Neural Network (ANN), Decision tree (C4.5), K-Nearest Neighbours and Support Vector Machine (SVM) | 20 features (not ratios) | Total—1000 samples |
| Heo and Yang (2014) | Adaboost with Decision Tree, SVM, Decision Tree, Artificial Neural Network | 12 ratios | 1381—Bankrupt firms<br>1381—non-bankrupt firms<br>Total—2762 firms |
| Du Jardin (2016) | Ensemble models, decision tree (DT), Multi-variable discriminant model (MDA), Logistic regression (LR) and Shallow Neural Network. | 35 ratios | 8010—Failed firms<br>8010—non-failed firms |
| Chen (2011) | Decision Tree, LDA, LR, Self-Organizing Map (SOM), Genetic algorithm, Learning vector optimization, Particle swarm optimization. | 8 features created using PCA over 42 ratios (33 financial and 8 non-financial). | 50—Bankrupt firms<br>150—non-bankrupt firms<br>Total—200 firms |
| Wang, Ma and Yang (2014) | FS-Boosting, LR, NB, DT, ANN, SVM, Ensemble models. | 30 financial ratios | 112—Bankrupt firms<br>128—non-bankrupt firms<br>Total—240 firms |
| Barboza, Kimura and Altman (2017) | Neural Network, Support Vector Machine with Linear kernel, Support Vector Machine with Radial Kernel, Boosting, Bagging and Random Forest | 11 financial ratios | Total—898 firms |

### *1.3. Significance of this study*

The previous studies done for bankruptcy prediction have not taken a systematic view of the data used to build the models. The previous studies have been more focused on the models rather than on the data

used to build the models. This study offers a much more balanced view where both the data and the models are given equal importance. To begin with, we have use Compustat are a source database to get an exhaustive list of financial ratios over US firms from 1970 to 2020. Compustat is a high-quality database used by several famous finance related papers such as Fama and French (1993). Most of the previous studies have used relatively small datasets as compared to ours. This study takes a systematic look at as many features as possible to train our machine learning models. Our balanced approach is also consistent with the shift from model centric to data centric approach proposed by Andrew Ng (Gil Press, 2021).

The rest of the paper is structured as follows:

Section 2—Describes the existing literature for bankruptcy prediction.

Section 3—Describes the data and the method used to clean, process, and fit the data into our machine learning models. This section also covers the process used to predict the number of bankruptcies using Q2-2020 ratios.

Section 4—Describes the results observed from the experiments

Section 5—Presents our final comments and discusses the implications of the results.

## 2. Literature review

Bankruptcy prediction models prior to 1990s were primarily statistical models employing univariate, multivariate and logit & probit techniques. In 1966, Beaver applied univariate analysis in which the predictive ability of 30 financial ratios was tested one at a time to predict bankruptcy (Beaver, 1966). Altman in 1968 performed a multi-variate discriminant analysis (MDA) using 5 ratios to create a linear discriminant function of 5 variables (Altman, 1968). Several variants of MDA were developed in the following years. Edmister used 19 financial ratios to build a linear model for bankruptcy prediction (Edmister, 1972). Deakin found that a linear combination of the 14 ratios could be used to predict bankruptcy five years prior to failure (Deakin, 1972). Ohlson studied the shortcomings of MDA models and built a conditional logit model using maximum likelihood estimation (Ohlson, 1980). The datasets used in all these studies were quite small as compared to modern standards. Ohlson's study for example used a dataset of 2058 firms out of which 105 firms represented the bankrupt class.

The next phase in the evolution of bankruptcy models started in the 1990s with several machine learning algorithms outperforming the older statistical models. Machine learning models such as Random Forests, Support Vector Machines (SVM) and Gradient Boosted Trees were found to be particularly effective for bankruptcy prediction. Barboza, Kimura and Altman compared statistical models with machine learning (ML) models. They found the Random Forests outperformed Alman's Z-score model by a significant margin (Barboza et al., 2017). These results were corroborated by studies (Joshi et al., 2018; Rustam and Saragih, 2018; Gnip and Drotár, 2019). Support Vector Machine (SVM) was also found to be a very effective machine learning algorithm in several studies. Hang et al. (2004) and Chen et al. (2008) achieved superior results for credit rating classification problem by using SVM. Song et al. (2008) used SVM to predict financial distress. Some studies also found boosted trees-based algorithms to outperform SVM. Wang, Ma and Yang proposed a new boosted tree-based algorithm for bankruptcy prediction which they found to be more effective than SVM (Wang et al., 2014). Heo and Yang (2014) used Adaboost algorithm to predict bankruptcy for Korean construction firm. They found Adaboost to have better accuracy than SVM (Heo and Yang, 2014). A more recent study in 2021 has used XGBoost and Random Forest

algorithms to predict bankruptcies over 12 months. This study used a medium sized training dataset containing data for 8959 firms registered in Italy (Perboli and Arabnezad, 2021). Another recent study uses a database of Taiwanese firms to predict bankruptcy. This study used data set contain 96 attributes for 6819 firms to train machine learning models (Wang and Liu, 2021). One common attribute shared by all the forementioned studies is the relatively small size of their training data sets. The datasets used by these studies are small as compared to datasets used in the big data era. The largest training dataset in these studies had just 2600 samples which is quite small.

Based on the literature review, the following trends become apparent:

- Machine Learning Models are now consistently outperforming statistical models
- The training data sets used to train the existing machine learning models are relatively small as compared to the data sets used for training models in other application areas.
- Ensemble methods such as Random Forest and Boosted trees have performed better than other models in bankruptcy prediction.

## 3. Data and methodology

This study differentiates itself from previous studies by using a substantially larger dataset as compared to previous studies. We use a very standard and well documented dataset called Compustat to retrieve the financial ratios. Compustat is a standard financial dataset used in financial research. Compustat has been used by some very popular papers in finance such as Fama and French (1993). We have used 57 financial ratios that are listed in Table 2. Financial ratios are inputs used to train Bankruptcy prediction models. While most studies use fewer financial ratios, this study applies a large set of financial ratios of US Firms from 1970–2020 (50 years) to train Random Forest, SVM and XGBoost Models. This section discusses the overall methodology which includes data cleaning, balancing, model fitting, and analysis of results.

### *3.1. Data*

Previous studies have used small to medium sized data sets for training Machine learning models. This study sets itself apart by using a much larger training dataset. We used financial ratios data set from Compustat. The financial ratios data set was then joined with another dataset called Bankruptcy data set. The bankruptcy data set contains the data such as date of bankruptcy, bankruptcy reason and GVKEY (primary key) while the financial ratios dataset contains all the financial ratios mentioned in Table A.1. The two datasets were programmatically joined using a common field named GVKEY. GVKEY is a unique identifier assigned to each firm. The relation amongst the two datasets that were used to create our labelled training dataset is best represented by the ER schema diagram shown in Figure 1.

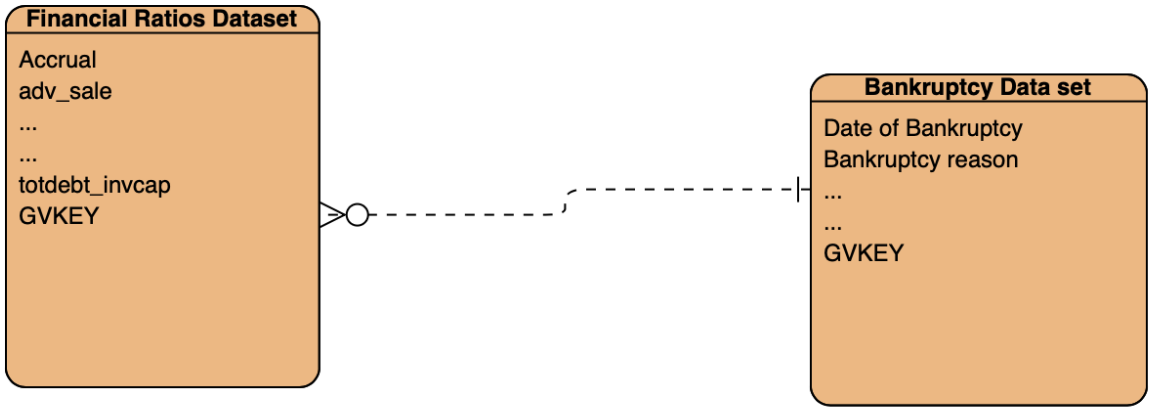


**Figure 1.** ER Diagram depicting relation between Financial Ratios Dataset and Bankruptcy Dataset.

The financial ratios dataset we have used contains 57 financial ratios mentioned in Table A.1 in Appendix A. This is an exhaustive list of features used to train our models. We have included ratios which are often overlooked but are likely to help detect patterns related to edge cases.

### *3.2. Methodology*

#### 3.2.1. Data preprocessing

The first step of building a predictive model is data pre-processing and cleaning. The original data from Compustat had 75 financial ratios for 21,114 US firms. This data covered firms established in the US between 1970 and 2020. The dataset contained firms that belonged to 2 classes: bankrupt and non-bankrupt or continuing enterprises. The dataset contains 1212 bankrupt firms and 19,902 non-bankrupt firms. The distribution of data points (samples) belonging to these two classes is summarized in Table 2. The next step was to drop features which had null values for more than 6000 firms out of 21,114 firms. This step ensured that we don't have more than 30% of null values in any feature. The goal is to ensure that the true distribution generating this data is preserved and learned by our machine learning models. 18 features (financial ratios) were dropped from the data set because they had null values for more than 6000 (30% of total number of firms). The dataset now had $75 - 18 = 57$ features. Next, we scaled our data to have mean = 0 and variance = 1 using Scikit-learns Standard Scaler class. Scaling is required to ensure that gradient descent converges on the minima of the loss function. The last step of data cleaning was to impute the missing values in the 57 financial ratios (features). For imputing the missing values, we used the KNN algorithm which used three nearest neighbours to estimate the missing value. Further, the weight assigned to each neighbour is a function of its Euclidean distance from the data point with missing value. KNN with 3-neighbours has been found to be effective in preserving the true distribution of the data (Beretta and Santaniello, 2016).

**Table 2.** Distribution of training data.

| Class | Count |
|---|---|
| Bankrupt | 1,212 |
| Non-Bankrupt | 19,902 |
| Total | 21,114 |

### 3.2.2. Balancing the dataset

The cleaned and scaled dataset without any missing values was an imbalanced dataset (see Table 2). The dominant class was the bankruptcy class. Approximately 90% of the samples belonged to the majority class which is non-bankrupt firms. Since the goal of this study is train a classifier to identify bankrupt firms, we decided to balance the classes in our training data. This would ensure that our model would learn about the minority class which is the bankrupt class. This is important in the context of bankruptcy prediction because detecting samples belonging to the bankrupt class. To balance the dataset, we use the Synthetic Minority Over-sampling technique (SMOTE) proposed by Chawla et al. (Chawla et al., 2002). SMOTE generates synthetic samples using the features of the data. The minority class is oversampled by taking a minority class sample and then a line is drawn from this minority class sample to k-nearest minority class samples. Synthetic minority class samples are generated along the line joining the minority class sample to its minority class neighbours. Additionally, to ensure that our balanced dataset facilitated learning of the bankrupt class, we also used Borderline-SVM SMOTE. Borderline-SVM SMOTE technique uses samples close the decision boundary (support vectors) to create synthetic samples (Nguyen et al., 2011). Finally, we used the Adaptive Synthetic Sampling (ADASYN) algorithm of He, Bai, Garcia and Li to generate samples in regions of feature space where the density of minority samples is low (He et al., 2008). The result was a balanced dataset containing 19902 samples of non-bankrupt class and 20,517 bankrupt class. The balanced dataset has 57 financial ratios (features).

### 3.2.3. Fitting training data into models

The balanced dataset was then shuffled and split into training set containing 70% of the samples and test set containing 30% of the samples. The purpose of creating a test set is to test the accuracy of the models on data that the models have not been trained on. Collecting metrics based on the test set gives practitioners an idea of the generalization performance of machine learning models.

The training data set was fitted into three machine learning models. These models are: Random Forest, Support Vector Machine (SVM) and XGBoost. After fitting, the models were then used to predict for samples in the test set to assess their relative performance.

### 3.2.4. Performance analysis

For comparing the performance of the models, we decided to use Accuracy score, Receiver Operating Curve (ROC) and Area Under ROC Curve (AUC). Accuracy score can be used because we are training our models using a balanced dataset. However, to get a better idea of the True Positive

Rate (TPR) and False Positive Rate (FPR) we decided to employ ROC and AUC metrics as well. It is important to compare the TPR and FPR because it is more to avoid False negatives (FN) as compared to False positives (FP). False negative (FN) would be a firm which would go bankrupt but is wrongly classified by our model as a non-bankrupt sample. False positive (FP) on the other hand would be a firm that is not bankrupt but is wrongly classified as a bankrupt firm.

### 3.2.5. Predicting bankruptcies

The goal of this study is to predict number of bankruptcies within the next 30, 90 and 180 days. We trained 3 different models to predict the number of bankruptcies within 30, 90 and 180 days. The models were built and analysed using the same approach. The only difference was that the training and test labels for each model were derived from the bankruptcy date. For example, to train the model for predicting number of bankruptcies within 30 days, we used

$$X = Matrix\ of\ shape\ (M, N)$$

*where*

$$M = Number\ of\ Firms\ (samples)$$

$$N = Number\ of\ features\ (Ratios)$$

$$Y = Vector\ of\ Shape\ (M, 1)$$

*where*

$$Y_i = 1\ if\ Firm\ i\ went\ bankrupt$$

*within* 30 *days of publishing financial ratios*

*else*

$$Y_i = 0$$

Therefore, we trained 9 models to predict bankruptcies within 30, 90 and 180 days. For example, for predicting bankruptcy within 30 days we trained Random Forest, SVM and XGBoost. After training the models, we picked the best model based on performance metrics described in previous section and then we used the best model to predict the number of bankruptcies using the latest Q2 2020 financial ratios from Capital IQ. In this final prediction set, we only kept data for firms which did not have any significant gaps or holes. Finally, we used the final prediction set from Q2 2020 to predict the number of bankruptcies we expect to happen over the next 30, 90 and 180 days.

## 4. Results

As mentioned in the previous section, we trained 9 models, using three different techniques, RF, SVM, XGBoost, for predicting bankruptcies over 30, 90 and 180 days. Next, we used the test set to make predictions and then assessed the relative performance. Based on the chosen metrics of accuracy score and Area under ROC curve (ROC AUC), XGBoost outperformed the other models for predicting bankruptcy within 30, 90 and 180 days. The actual scores for accuracy and AUC are presented in Table 3.

**Table 3.** Accuracy and ROC AUC metrics.

| Model | Algorithm | Accuracy Score | ROC AUC |
|---|---|---|---|
| Predict Bankruptcies within 30 days | Random Forest (RF) | 0.99676 | 0.99981 |
| | Support Vector Machine | 0.90933 | 0.96673 |
| | XGBoost | 0.99683 | 0.99992 |
| Predict Bankruptcies within 90 days | Random Forest (RF) | 0.98654 | 0.99917 |
| | Support Vector Machine | 0.83528 | 0.90526 |
| | XGBoost | 0.99047 | 0.99933 |
| Predict Bankruptcies within 180 days | Random Forest (RF) | 0.98580 | 0.99896 |
| | Support Vector Machine | 0.82202 | 0.89590 |
| | XGBoost | 0.98697 | 0.99902 |

The accuracy score of XGBoost models is consistently better than SVM and Random Forest. This result is also consistent with the ROC curves which are shown in Figure 2 below.

As seen in Figure 2, the ROC curve for XGBoost is closest to the top left corner thereby covering maximum area under it. XGBoost is therefore the best performing model closely followed by Random Forest. The fact that these metrics are calculated using the test set (containing data which model has not been trained on) gives us confidence in the ability of our models to generalize.

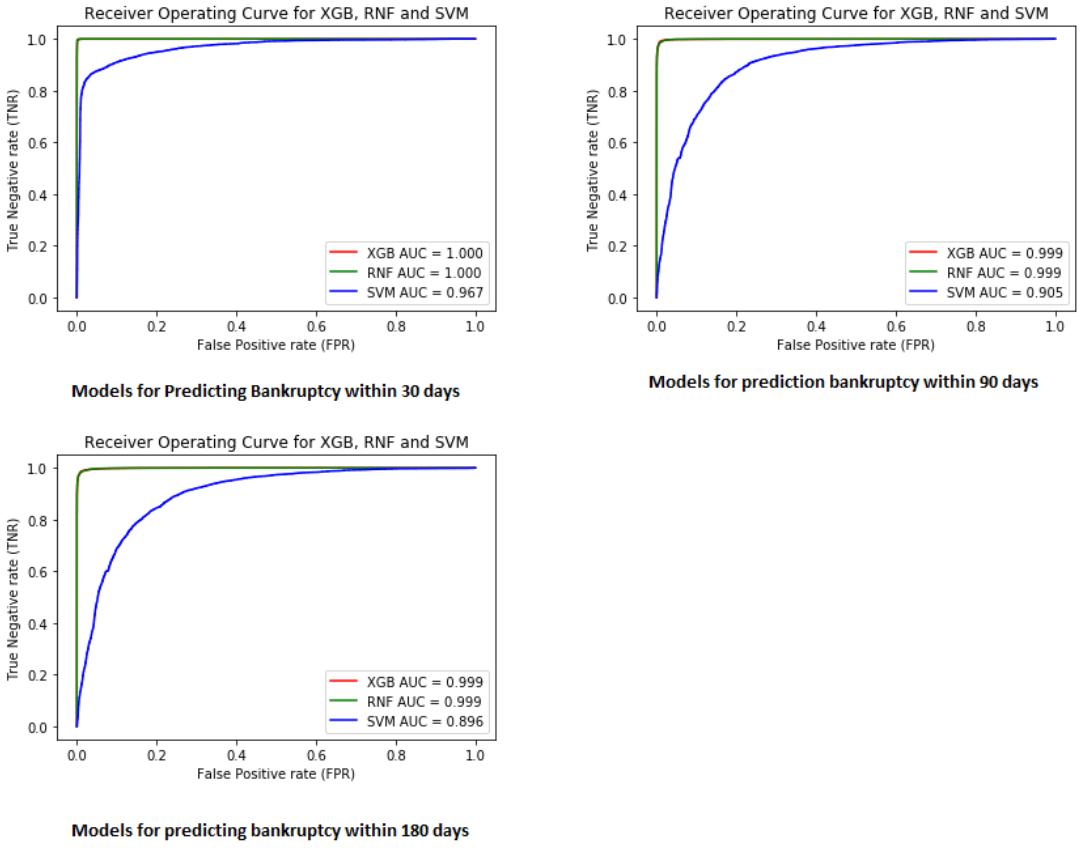


**Figure 2.** ROC curves for all 9 models.

### *4.1. Comparison of performance with previous studies*

We present the performance metrics of previous studies in Table 4. Previous studies have used 2 performance metrics: Test accuracy and Area Under ROC curve (AUC). To keep the comparison consistent, we have computed both test accuracy score and AUC for our models (see Table 3). Our best model built using XGBoost significantly outperforms the models built in previous studies. The accuracy of our XGBoost model for prediction bankruptcy within 180 days is 98.69% which is lower than the test accuracy of our XGBoost models for predicting bankruptcies within 30 and 90 days. However, our model for predicting bankruptcies within 180 days has a higher test accuracy (98.69%) than models built in previous studies. Similarly, our model for predicting bankruptcies within 180 days has an AUC score of 0.99 which is higher than the AUC score reported by previous studies. Our performance metrics of accuracy and AUC score are computed over out of training samples which also indicates to the robustness of our results.

### *4.2. Predicting bankruptcies caused by Covid*

Next, we apply our best model, using XGBoost, to the data from Q2-2020 to evaluate the possibility of a substantial upsurge in business bankruptcies in the second half of 2020 because of the deep 2020 recession caused by the pandemic. We apply this best model to the latest available ratios, for Q2-2020, and classify a firm as going bankrupt during the next 30, 90, or 180 days if the predicted probability of bankruptcy is higher than 0.50.

Using this method, our best model in each category predicted 74 bankruptcies within 30 days, 189 bankruptcies within 90 days and 354 Bankruptcies within 180 days. This prediction is for all firms contained in the S&P Global database, both public and private. The predictions for the number of bankruptcies are summarized in Table 5.

S&P Global has reported a total of 336 actual bankruptcies until the end of June 2020. If we add our prediction of 354 bankruptcies to the actual bankruptcies, then we predict a total of 336 + 354 = 690 bankruptcies in 2020. We summarize our predictions in Table 6 below.

Since the number of firms in the database changes from year to year, we decided to compare the prediction for 2020 with the past by using bankruptcy rates, i.e., the ratio of the number of bankruptcies to the total number of firms. As shown in Table 7, our prediction of 690 bankruptcies in 2020 represents a bankruptcy rate of 4.35% for all US firms. This rate is the highest in the last 10 years. The second highest rate of 4.2%, only slightly lower, was seen in 2010, in the immediate aftermath of the 2008-09 recession. The average rate during the economic expansion years of 2011–2019 was 3.2%, more than a full percentage point lower than the predicted 2020 rate. We conclude that we will indeed see a much higher rate bankruptcies in 2020, but it is unlikely to be substantially larger than in 2010.

**Table 4.** Performance metrics of previous studies.

| Author | Size of training set | Performance |
|---|---|---|
| Wilson and Sharda (1994) | 65—Bankrupt firms | Test set accuracy* |
| | 64—non-bankrupt firms | 95.6% for neural network |
| | Total—169 firms | 91.8% for MDA |
| Min and Lee (2005) | 944—Bankrupt firms | Test set accuracy* |
| | 944—non-bankrupt firms | 83.06% for SVM |
| | Total—1888 firms | 82.52% for Shallow Neural Network |
| | | 79.13% for MDA |
| | | 78.30% for Logit |
| Fedorova, Gilenko and | 444—Bankrupt firms | Test set accuracy* |
| Dovzhenko (2013) | 444—non-bankrupt firms | 88.8% for Adaptive Boost combining |
| | Total—888 firms | several NN |
| | | 87.8% for Logistic regression |
| Shi, Xi, Ma, Hu (2009) | Total—1000 samples | Test set accuracy* |
| | | 75.6% for Bagging ensemble of ANNs |
| Heo and Yang (2014) | 1381—Bankrupt firms | Test set accuracy* |
| | 1381—non-bankrupt firms | 78.5% for AdaBoost |
| | Total—2762 firms | 77.1% for ANN |
| | | 73.3% for SVM |
| Du Jardin (2016) | 8010—bankrupt firms | Area under ROC Curve (AUC)** |
| | 8010—non-bankrupt firms | 0.9049 for Neural Network with Random |
| | | subspace ensemble |
| | | 0.9003 for Neural Networks with Boosting |
| | | 0.8952 for Neural Networks with Bagging |
| Chen (2011) | 50—Bankrupt firms | Test set accuracy* |
| | 150—non-bankrupt firms | 93.12% for PSO-SVM |
| | Total—200 firms | 91.87% for GA-SVM |
| | | 84.37 for SVM |
| Wang, Ma and Yang (2014) | 112—Bankrupt firms | Test set accuracy* |
| | 128—non-bankrupt firms | 81.50% for FS-Boosting |
| | Total—240 firms | 72.21% for SVM |
| | | 73.38% for ANN |
| Barboza, Kimura and Altman | 449—Bankrupt firms | Area under ROC curve (AUC) ** |
| (2017) | 449—non-bankrupt firms | Random Forest—92.92 (highest AUC). |
| | Total—898 firms | |

Note: *Studies that reported test set accuracy

**Studies that reported Area under ROC curve

**Table 5.** Bankruptcy predictions using Q2 2020 data.

| Model | Predicted Bankruptcies |
|---|---|
| XGBoost for Predicting Bankruptcies within 30 days | 74 |
| XGBoost for Predicting Bankruptcies within 90 days | 189 |
| XGBoost for Predicting Bankruptcies within 180 days | 354 |

**Table 6.** Total number of bankruptcy predictions.

| Time Period | Bankruptcies |
|---|---|
| Actual bankruptcies reported until Jun-2020 | 336 |
| Predicted bankruptcies from our model, Jul-Dec 2020 | 354 |
| Total bankruptcies for the year 2020 | 690 |

**Table 7.** Comparison of bankruptcy rate with past bankruptcy rates.

| Year | # Of Bankruptcies | # Of Firms | % Of Bankruptcies |
|---|---|---|---|
| 2010 | 819 | 19,523 | 4.20% |
| 2011 | 629 | 19,001 | 3.31% |
| 2012 | 582 | 18,653 | 3.12% |
| 2013 | 551 | 18,373 | 3.00% |
| 2014 | 467 | 18,091 | 2.58% |
| 2015 | 520 | 17,181 | 3.03% |
| 2016 | 571 | 17,004 | 3.36% |
| 2017 | 513 | 17,118 | 3.00% |
| 2018 | 513 | 16,542 | 3.10% |
| 2019 | 582 | 14,442 | 4.03% |
| 2020 | 690 | 15,850 | 4.35% |

## 5. Conclusions

We find that two different Machine Learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) produce accurate predictions of whether a firm will go bankrupt within the next 30, 90, or 180 days, using financial ratios as input features. The XGBoost based models perform exceptionally well, with 99% out-of-sample accuracy. Our training dataset uses a large database of public US firms over a period of 49 years, 1970–2019, and 57 financial ratios. This study has used a substantially larger training dataset as compared to previous studies.

An application of our best performing XGBoost model to Q2-2020 financial data for a sample of both private and public U.S. firms shows that the bankruptcy rate will climb substantially higher in 2020 than in the expansion years of 2011–2019. However, our model suggests that the rate will be only marginally higher than in 2010.

### *5.1. Avenues for future research*

We identify the following areas for further research:

- Adding macro-economic features—It will be interesting to add macro-economic features to training data used for training machine learning models for bankruptcy prediction.
- Train deep neural networks with different topologies—Another interesting area of research would be to apply different types of deep neural networks such as TabNet and Recurrent neural networks.

## Conflict of interest

All authors declare no conflicts of interest in this paper.

## References


Altman EI (1968) The Prediction of Corporate Bankruptcy: A Discriminant Analysis. *J Financ* 23: 193–194.

Barboza F, Kimura H, Altman E (2017) Machine learning models and bankruptcy prediction. *Expert Syst Appl* 83: 405–417.

Beaver WH (1966) Financial Ratios As Predictors of Failure. *J Account Res* 4: 71–111.

Bellovary JL, Giacomino DE, Akers MD (2007) A Review of Bankruptcy Prediction Studies: 1930 to Present. *J Financ Educ* 33: 1–42.

Beretta L, Santaniello A (2016) Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inf Decis Making* 16: 197–208.

Cao B, Zhan D, Wu X (2009) Application of svm in financial research, In: *CSO international joint conference on computational sciences and optimization*, 2: 507–511.

Chava S, Jarrow RA (2004) Bankruptcy Prediction with Industry Effects. *Rev Financ* 8: 537–569.

Chawla NV, Bowyer KW, Hall LO, et al. (2002) SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 16: 321–357.

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system, In: *Proc. of KDD'16*, 785–794.

Chen WM, Ma CQ, Feng GB (2008) Application of SVM based on KMOD function in credit scoring. *Math Econ* 25: 24–27.

Chen MY (2011) Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Comput Math Appl* 62: 4514–4524.

Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20: 273–297.

Deakin EB (1972) A Discriminant Analysis of Predictors of Business Failure. *J Account Res* 10: 167–179.

Du Jardin P (2016) A two-stage classification technique for bankruptcy prediction. *Eur J Oper Res* 254: 236–252.

Edmister RO (1972) An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction. *J Financ Quant Anal* 7: 1477–1493.

Fama, EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *J Financ Econ* 33: 3–56.

Fedorova E, Gilenko E, Dovzhenko S (2013) Bankruptcy prediction for Russian companies: Application of combined classifiers. *Expert Syst Appl* 40: 7285–7293.

García V, Marqués AI, Sánchez JS (2015) An insight into the experimental design for credit risk and corporate bankruptcy prediction systems. *J Intell Inf Syst* 44: 159–189.

Gil Press (2021) Andrew Ng Launches A Campaign For Data-Centric AI. Forbes. Available from: https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/?sh=1b802f8d74f5.

Gnip P, Drotár P (2019) Ensemble methods for strongly imbalanced data: bankruptcy prediction, In: *2019 IEEE 17th International Symposium on Intelligent Systems and Informatics (SISY)*, 155–160.

He H, Bai Y, Garcia EA, et al. (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning, In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328.

Heo J, Yang JY (2014) AdaBoost based bankruptcy forecasting of Korean construction companies. *Appl Soft Comput* 24: 494–499.

Huang W, Nakamori Y, Wang SY (2005) Forecasting stock market movement direction with support vector machine. *Comput Oper Res* 32: 2513–2522.

Huang Z, Chen H, Hsu CJ, et al. (2004) Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decis Support Syst* 37: 543–558.

Joshi S, Ramesh R, Tahsildar S (2018) A Bankruptcy Prediction Model Using Random Forest, In: *Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1–6.

Lyandres E, Zhdanov A (2007) Investment Opportunities and Bankruptcy Prediction. SSRN Electronic Journal. Available from: https://doi.org/10.2139/ssrn.946240.

Min JH, Lee YC (2005) Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Syst Appl* 28: 603–614.

Nguyen HM, Cooper EW, Kamei K (2011) Borderline over-sampling for imbalanced data classification. *Int J Knowl Eng Soft Data Paradigms* 3: 4–21.

Ohlson JA (1980) Financial Ratios and the Probabilistic Prediction of Bankruptcy. *J Account Res* 18: 109–131.

Perboli G, Arabnezhad E (2021) A Machine Learning-based DSS for mid and long-term company crisis prediction. *Expert Syst Appl* 174: 114758.

Rustam Z, Saragih GS (2018) Predicting Bank Financial Failures using Random Forest. In *2018 International Workshop on Big Data and Information Security (IWBIS)*, 81–86.

Shi L, Xi L, Ma X, et al. (2009) Bagging of Artificial Neural Networks for Bankruptcy Prediction. *2009 International Conference on Information and Financial Engineering*, Singapore, 154–156.

Song JK, Zhang ZX, Zhang Y (2008) Financial distress early-warning of companies based on multiclassification SVM. *China Manage* 4: 47–49.

Wang G, Ma J, Yang S (2014) An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Syst Appl* 41: 2353–2361.

Wang H, Liu X (2021) Undersampling bankruptcy prediction: Taiwan bankruptcy data. *PLoS ONE* 16: e0254030.

Wilson RL, Sharda R (1994) Bankruptcy prediction using neural networks. *Decis Support Syst* 11: 545–557.