VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
UNIVERSITY OF ECONOMICS AND LAW

BANKRUPTCY PREDICTION OF POLISH ENTERPRISES BY MACHINE LEARNING BASED ON FINANCIAL RATIOS

SUBJECT: MACHINE LEARNING

LECTURER: Assoc. Prof. Nguyen Anh Phong
MFin. Phan Huy Tam

Submitted by: Pham Quynh Huong
ID: K194141724

JUNE 2022

TABLE OF CONTENTS

1. Introduction

1.1. Motivation

Estimating the risk of corporate bankruptcies is of large importance to creditors and investors. There are large indirect and direct factors associated with financial distress and bankruptcies and for this reason bankruptcy prediction has for a long period of time constituted an extensive area of research all over the world. Corporate bankruptcies can have serious effects both locally and globally, employees, investors, customers, suppliers and their financiers are all affected when a company disappears. More serious, in some cases, a corporate bankruptcy can cause an entire industry to suffer. Until recently the dominating methods for predicting corporate bankruptcies have been based on statistical modeling, however, lately models based on machine learning have been proposed. Recently, machine learning models have successfully been used for many classification and regression problems and these models have often outperformed traditional methods for prediction purposes. The purpose of bankruptcy prediction is to assess the financial health status and future perspectives of a company. From there, we can warn investors earlier, creditors about the bankruptcy risk of companies. Moreover, it helps managers have a convenient technical way to assess a firm's operating performance and financial risk.

1.2. Literature review

Initial bankruptcy prediction models were primarily statistical models employing univariate, multivariate. 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). On the other hand, 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). 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 with several Machine Learning algorithms outperforming the older statistical models. Machine Learning models such as Random Forests, Support Vector Machines and Gradient Boosted Trees were found to be particularly effective for bankruptcy prediction. Barboza, Kimura and Altman compared statistical models with Machine Learning models. They found the Random Forests outperformed Alman's Z-score model by a significant margin (Barboza et al., 2017). Support Vector Machine was also found to be a very effective machine learning algorithm in several studies. Song et al. (2008) used SVM to predict financial distress. 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 a data set containing 96 attributes for 6819 firms to train Machine Learning models (Wang and Liu, 2021).

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

(1) Machine Learning Models are now consistently outperforming statistical models in the financial sector.
(2) The efficient application of Machine Learning in prediction for classification finance models.
(3) Ensemble methods such as Random Forest and Boosted trees have performed better than other models in bankruptcy prediction.

1.3. Objectives of the study

The intention of the study is to illustrate and investigate how machine learning can be exploited for prediction of corporate bankruptcies. The fact is that Vietnam is a developing economy and has very

few firms which were bankrupt before, so there is a lack of sources providing Vietnam's bankrupt data set. That is why we have decided to research the data set from Poland and have expected to gain experience for the Vietnam situation after. In our study we use financial ratios to address bankruptcy prediction of Polish companies in the manufacturing sector between 2000 and 2013. The intention of this study is to produce, predict and evaluate the effect of the application Machine Learning, specifically Random Forest model, for the bankruptcy possibility based on financial ratio consideration. Therefore, this research will be expected to propose a Random Forest performance on the Polish data set. Moreover, after constructing a model, we can identify key success factors for the performance of Random Forest in the context of corporate bankruptcy prediction. Lastly, we want to present at which time we can work on the most correct prediction in the period of 5 years until bankruptcy.

2. Theory

2.1 Introduction to machine learning

2.1.1 Definition of Machine learning

Machine Learning, according to Arthur Samuel, is "the branch of research that makes computers capable of learning without being explicitly programmed."

Machine Learning, as defined by Tom Mitchell, is "a computer program that learns from experience E to do task T, and its effectiveness is evaluated by P, if its efficacy in executing task T is measured by performance P, enhanced by experience E."

In this study, T E P is defined as:

(1) Task T is to determine whether the business is profitable or not.
(2) Experience E is the characteristic to classify businesses that profit from available data.
(3) The performance P metric is the accuracy of the determination process.

Machine learning is the activity of learning from data in an iterative fashion using various algorithms to develop models and predict outcomes. The data set is divided into two pieces by machine learning: the training set and the test set. Algorithms employ training data to build machine learning models. The test data set will be used to assess the correctness of the produced model.

Popular machine learning algorithms such as: Artificial Neural Networks - ANN, Support Vector Machines - SVM, Genetic Programming - GPN, K-nearest neighbors (KNN), Logistic regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest…

2.1.2. Reasons for choosing Machine learning

Both Machine Learning and Statistical Estimation techniques can generate forecasts. However, each type has different strengths and purposes. In it, machine learning models are designed to make the most accurate predictions possible. Statistical models are designed to make inferences about relationships between variables.

Statistics is the mathematical study of data. You can't make statistics unless you have the data. A statistical model is a model for data that is used to infer something about relationships within the data or to create a model that can predict future values. Therefore, there are many statistical models that can make predictions, but the accuracy of the predictions is not their strong point.

In contrast, Machine Learning models provide varying degrees of interpretability, from highly interpretable Lasso Regression to impenetrable neural networks, but they often sacrifice interpretability for predictability. The purpose of machine learning is to obtain a model that can make reproducible predictions without regard to whether the model is interpretable or not, although you should still experiment to make sure. ensure that the model's predictions make sense.

So, which method is better depends on what you use it for. If you just want to create an algorithm that can predict house prices with great accuracy, or use data to determine if someone is likely to have certain types of disease, then machine learning might be the better approach. If you are trying to prove relationships between variables or make inferences from data, statistical modeling may be a better approach.

In this study, our goal is to be able to most accurately predict which businesses are profitable. Therefore, using Machine Learning will be a more optimal method and also a newer method than previous studies using regression statistics to predict results.

Econometrics is mainly concerned with model interpretation, which is the evaluation of the effects of independent variables on a large number of dependent variables in an econometric model. Machine learning is exclusively concerned with the model's prediction results; ML models attempt and fail to find the best accurate prediction. Machine learning is better suited to prediction than econometrics.

2.1.3. Supervised learning

Supervised learning is a method of Machine Learning that uses labeled data with the main goal of determining the relationship between input and output variables. Supervised learning is divided into two problems: regression and classification. Regression is used to predict continuous data, while classification is used for discrete data.

2.1.4. Random forest

According to the definition of Breiman (2001): random forest is “a classifier consisting of a collection of tree-structured classifiers {h(x, Θk), k = 1, . . .} where the {k} are independently identically distributed random vectors and each tree casts a unit vote for the most popular class at input x”.

The random forest classification algorithm is an enhanced version of the decision tree classification technique. This algorithm, which is built from several decision trees, aids in overcoming the problem of overfitting. To pick trees in the forest, the algorithm utilizes a voting mechanism.

There are two ways to vote on a random forest. One is to choose the decision tree with the highest number of votes. The second is to select the results based on the proportion of votes, these votes are the weight of the results.

Random forest generates random trees by: bootstrapping and criteria selection.

(1) Bootstrapping technique: Each tree is created with a unique data set that is made up of a subset of the same size of the available data.
(2) Criteria: Decision trees will choose features for the tree to branch.

The random forest can indicate the importance of the model's features. It helps us to know the most important features and the unimportant features. This helps to focus more on the essentials and can be considered to remove low-impact features.

2.2 Measurement metrics

2.2.1 Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. It is defined as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ predictions}$$

A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified. For example, if the algorithm classified a false data point as true, it would be a false positive. Often, accuracy is used along with precision and recall, which are offered before. Together, these metrics provide a detailed look at how the algorithm is classifying data points.

2.2.2 Confusion matrix

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Some features of Confusion matrix are given below:

(1) For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.

(2) The matrix is divided into two dimensions, that are predicted values and actual values along with the total number of predictions. Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

n = total predictions	**Actual negative**	**Actual positive**
Predicted negative	TN	FN
Predicted positive	FP	TP

Table 1. The sample confusion matrix

(1) **TN** (True Negative): Model has given prediction No, and the real or actual value was also No.

(2) **TP** (True Positive): The model has predicted yes, and the actual value was also true.

(3) **FN** (False Negative**)**: The model has predicted no, but the actual value was Yes, it is also called a Type-II error.

(4) **FP** (False Positive): The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

2.2.3 Sensitivity and specificity

(1) Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive). Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed as the false negative). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1. Mathematically, sensitivity can be calculated as the following:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

The higher value of sensitivity would mean higher value of true positive and lower value of false negative. The lower value of sensitivity would mean lower value of true positive and higher value of false negative. For the healthcare and financial domain, models with high sensitivity will be desired.

(2) Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

The higher value of specificity would mean higher value of true negative and lower false positive rate. The lower value of specificity would mean lower value of true negative and higher value of false positive.

2.2.4 Receiver operating characteristic

ROC Curve stands for Receiver Operating Characteristics Curve, which is a metric to evaluate the performance of a classification model. Each point along the ROC Curve corresponds to a classification model. The ROC curve shows the sensitivity of the classification model to the ratio between true positive and false negative. The ROC that compares two operating characteristics as the criteria change is called a relative operating characteristic curve. These two performance characteristics are True Positive Rate (TPR) and False Positive Rate (FPR). An ROC with a ratio between True Positive and False Positive corresponding to 1:0 is considered ideal.

Important points in ROC Curve:

(1) TPR = 0, FPR = 1: the model predicts all cases to be negative class
(2) TPR = 1, FPR = 1: The model predicts all cases to be positive class
(3) TPR = 1, FPR = 0: Ideal model with 0 false classifications.

2.2.5 Area under curve

AUC, also known as Area Under Curve, is a metric commonly used to evaluate machine learning models, which is the area below the ROC curve. The AUC helps the classifier to distinguish between classes. The higher the Area Under Curve, the better the positive and negative discrimination performance of the model.

The classifier is considered perfect when AUC = 1. And AUC=0 if the algorithm only makes random guesses.

3. Data

3.1. Data resources

3.1.1. Data source

The data set used in this study consisted of financial information about Polish companies in the manufacturing sector. The data set contained information about both bankrupt companies and still operating ones. The financial information was extracted from the database Emerging Markets Information Service (EMIS). The financial indicators describing the health of the bankrupt companies were collected during 2007-2013 and the information about the still operating companies was gathered between 2000-2012. For simplicity, we will refer to the companies that went bankrupt as belonging to Class 1 and the surviving companies as belonging to Class 0.

The data is divided into five different subsets, which are described in table 2. The subsets are created to allow for different lengths of forecasting period. The task in each of them is to predict whether or not a company goes bankrupt within five, four, three, two and one years respectively based on information that could be retrieved from 64 different financial indicators. All five data sets are, as is expected also in real life, heavily imbalanced. There are much fewer bankrupt companies compared to still operating ones.
64 different financial indicators are used as ratios, so it means that the predictors are not too heavily correlated with the size of the companies. A complete list of the features is found in table 3.

Data set	Features from	Bankruptcy after	No. bankrupt	No. not bankrupt	Sum
first year	1_year	5 years	271	6756	7027
second year	2_year	4 years	400	9773	10173
third year	3_year	3 years	495	10008	10503
fourth year	4_year	2 years	515	9277	9792
fifth year	5_year	1 years	410	5500	5910

Table 2. The difference between five data subsets

ID	**Description**	**ID**	**Description**
Attr1	net profit / total assets	Attr33	operating expenses / short-term liabilities
Attr2	total liabilities / total assets	Attr34	operating expenses / total liabilities
Attr3	working capital / total assets	Attr35	profit on sales / total assets
Attr4	current assets / short-term liabilities	Attr36	total sales / total assets
Attr5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	Attr37	(current assets - inventories) / long-term liabilities
Attr6	retained earnings / total assets	Attr38	constant capital / total assets
Attr7	EBIT / total assets	Attr39	profit on sales / sales
Attr8	book value of equity / total liabilities	Attr40	(current assets - inventory - receivables) / short-term liabilities
Attr9	sales / total assets	Attr41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr10	equity / total assets	Attr42	profit on operating activities / sales
Attr11	(gross profit + extraordinary items + financial expenses) / total assets	Attr43	rotation receivables + inventory turnover in days
Attr12	gross profit / short-term liabilities	Attr44	(receivables * 365) / sales
Attr13	(gross profit + depreciation) / sales	Attr45	net profit / inventory

Attr14	(gross profit + interest) / total assets	Attr46	(current assets - inventory) / short-term liabilities
Attr15	(total liabilities * 365) / (gross profit + depreciation)	Attr47	(inventory * 365) / cost of products sold
Attr16	(gross profit + depreciation) / total liabilities	Attr48	EBITDA (profit on operating activities - depreciation) / total assets
Attr17	total assets / total liabilities	Attr49	EBITDA (profit on operating activities - depreciation) / sales
Attr18	gross profit / total assets	Attr50	current assets / total liabilities
Attr19	gross profit / sales	Attr51	short-term liabilities / total assets
Attr20	(inventory * 365) / sales	Attr52	(short-term liabilities * 365) / cost of products sold)
Attr21	sales growth rate	Attr53	equity / fixed assets
Attr22	profit on operating activities / total assets	Attr54	constant capital / fixed assets
Attr23	net profit / sales	Attr55	working capital
Attr24	gross profit (in 3 years) / total assets	Attr56	(sales - cost of products sold) / sales
Attr25	(equity - share capital) / total assets	Attr57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr26	(net profit + depreciation) / total liabilities	Attr58	total costs /total sales
Attr27	profit on operating activities / financial expenses	Attr59	long-term liabilities / equity
Attr28	working capital / fixed assets	Attr60	sales / inventory
Attr29	logarithm of total assets	Attr61	sales / receivables
Attr30	(total liabilities - cash) / sales	Attr62	(short-term liabilities *365) / sales
Attr31	(gross profit + interest) / sales	Attr63	sales / short-term liabilities
Attr32	(current liabilities * 365) / cost of products sold	Attr64	sales / fixed assets

Table 3. A complete table of the features

3.1.2. Research subjects

The data set used in this study consisted of financial information about Polish companies in the manufacturing sector. The data set contained information about both bankrupt companies in a period from 1 to 5 years and still operating ones. There are a total of 43405 companies in those 5 subsets.

3.1.3. Research scope

The financial indicators describing the health of the bankrupt companies were collected during 2007-2013 and the information about the still operating companies was gathered between 2000-2012.

3.1.4. Variables

(1) Target variable - "class" column contains 2 unique values which are 0 (no bankrupt) and 1 (bankrupt).
(2) Independent variables - 64 independent variables are described as table 3 above.

3.2. Preprocessing data

3.2.1. Drop high-correlation coefficient features

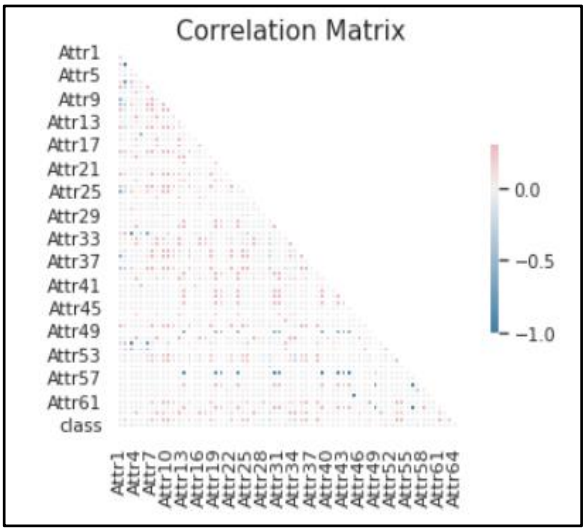


Figure 1. Correlation matrix of data set 1.

The correlation coefficients range from -100% to zero. That shows the data set has high correlations between various variables. Because there are several variables which have high correlation with each other, we conduct to drop the variables correlating more than 50% with from 2 other ones to up. Then, we practice similar to the 4 remaining data sets.

3.2.2. Drop columns which contain more than 5% missing values of total

Figure 2. Missing data map of data set 1

According to the figure above, there are a few columns containing many missing values. Dropping observations containing missing values will indirectly make this data set lose actual bankrupt companies. Because this data set is an imbalanced data set and the number of collapsing companies is very small compared with the rest others. So that we choose the optimal solution is dropping the columns which contain more than 5% missing values of total. The 4 remaining subsets are conducted similarly.

3.2.3. Fill remaining missing values with mean values

The remaining missing values are filled by their column's mean values. The reason for this choice is that the subset type is not a time series, so we can not fill by regression or backfill, forward fill. Moreover, all of the values in the subset are financial ratios which are not affected by company size, except for industry.

3.2.4. Merge 5 subsets into a big data set

After dealing with missing values and high correlation problems, we finally have 28 features including 27 independent variables and a target variable - "bankrupt". We will merge 5 subsets into a big data set. But before merging, we create an additional column named "amt_year", it means "the amount of year until bankruptcy" for distinguishing the amount of year until bankruptcy from each subset.

(1) The "df1" subset has a period of 5 years until bankruptcy.
(2) The "df2" subset has a period of 4 years until bankruptcy.
(3) The "df3" subset has a period of 3 years until bankruptcy.
(4) The "df4" subset has a period of 2 years until bankruptcy.
(5) The "df5" subset has a period of 1 years until bankruptcy.

Therefore, the final big data set has 29 columns consisting of 28 independent variables and a target variable with a total of 43405 observations.
The final independent variables are described at the table below:

ID	Description	ID	Description
Attr4	current assets / short-term liabilities	**Attr33**	operating expenses / short-term liabilities
Attr5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	**Attr34**	operating expenses / total liabilities
Attr8	book value of equity / total liabilities	**Attr40**	(current assets - inventory - receivables) / short-term liabilities
Attr9	sales / total assets	**Attr41**	total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr12	gross profit / short-term liabilities	**Attr46**	(current assets - inventory) / short-term liabilities
Attr15	(total liabilities * 365) / (gross profit + depreciation)	**Attr47**	(inventory * 365) / cost of products sold
Attr16	(gross profit + depreciation) / total liabilities	**Attr50**	current assets / total liabilities
Attr17	total assets / total liabilities	**Attr52**	(short-term liabilities * 365) / cost of products sold)
Attr26	(net profit + depreciation) / total liabilities	**Attr53**	equity / fixed assets
Attr28	working capital / fixed assets	**Attr54**	constant capital / fixed assets
Attr29	logarithm of total assets	**Attr55**	working capital

Attr63	sales / short-term liabilities	**Attr57**	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr64	sales / fixed assets	**Attr59**	long-term liabilities / equity
amt_year	the amount of years until bankrupt	**Attr61**	sales / receivables

Table 4. The final independent variables in the big data set

3.3. Visualize data

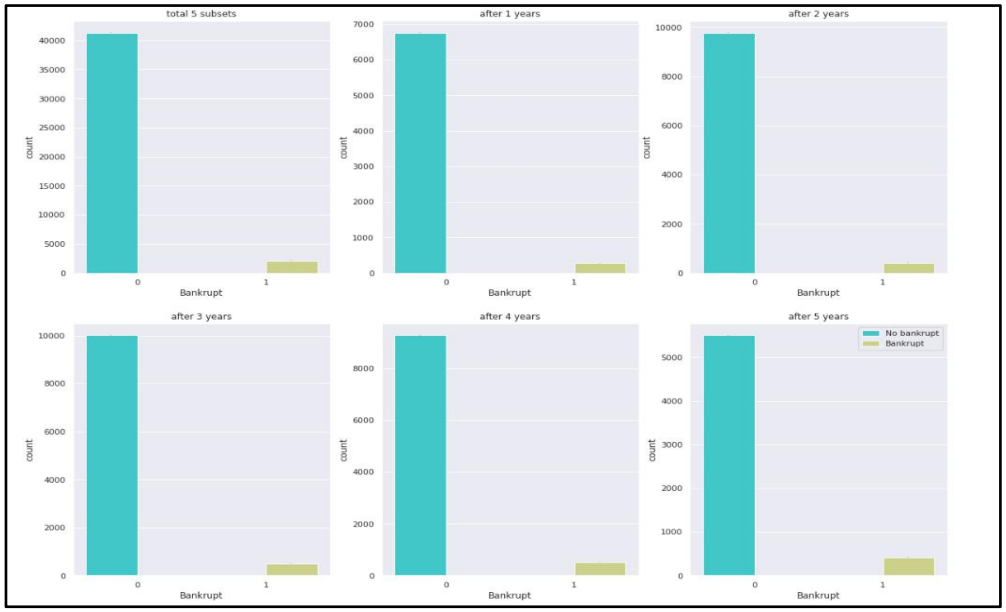


Figure 3. Visualization of bankruptcy distribution in the big data set and the 5 subsets

The six graphs above show the difference between the number of companies that will actually go bankrupt and not go bankrupt after 1 to 5 years, and aggregate those 5 cases. Through visualization, we can see that this dataset is unbalanced. During model building and reading results, the imbalanced dataset problem will be demonstrated more clearly.

4. Model and result

4.1. Choose the best model

4.1.1 Choose model for prediction

We build models on python3 programming language combined with open source software packages.

To construct the machine learning model, divide the data into two parts, a training set and a testing set with a ratio of 80:20. The training set is used to train the machine learning model, and the testing set is used to test the model. We test machine learning algorithms to find the most optimal algorithm: Logistic Regression, K-nearest neighbors, Decision above, Support vector Machine (Linear and RBF Kernel), Neural Network, and Random forest.

Algorithm	**Model score**
Logistic Regression	94.92%
K-Nearest Neighbors	94.70%

Decision Tree	92.55%
Support Vector Machine (Linear Kernel)	94.91%
Support Vector Machine (RBF Kernel)	94.90%
Neural Network	94.78%
Random Forest	94.95%

Table 5. Table of various models' scores in order to search for the best one

The score shows that Random forest is the best model to forecast corporate profits with 94.95%. According to various literature reviews and the models' scores table above, we decided to get Random Forest to be a main algorithm in this research.

4.1.2. Choose the best number of decision trees for Random Forest model

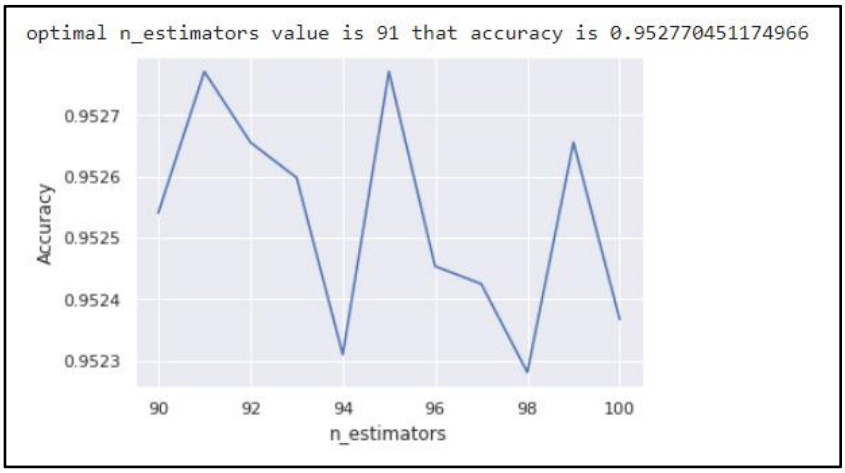
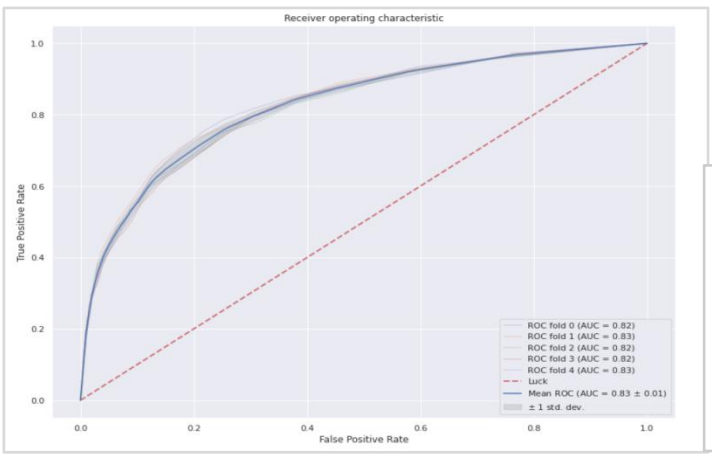


Figure 4. The GridSearch from 90 to 100 decision trees in Random Forest model

We perform GridSearch from 90 to 100 random forest to find the optimal number of trees. The results show that the model stopping at the 91st tree is optimal. We train the random forest model with the 91st optimal number of trees.

4.3. Result

4.3.1. ROC - AUC chart

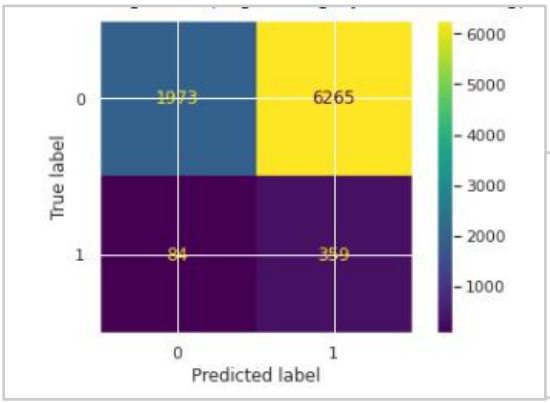


	AUC Train	AUC Test
0	1.000000	0.821927
1	0.999999	0.834304
2	0.999999	0.820912
3	1.000000	0.822729
4	1.000000	0.833524

Figure 5. ROC chart and train-test AUC scores of the original data set

Also, the ROC-AUC findings of this model are pretty good, as shown in Figure 5. The 5-fold AUC values vary from 0.82 - 0.83, showing that the model is doing well. It does not have any overfitting problem in this model, because the difference between training AUC scores and test AUC scores is quite low after running 5 times with KFold function.

4.3.2. Classification report



	precision	recall	f1-score	support
0	0.96	0.24	0.38	8238
1	0.05	0.81	0.10	443
accuracy			0.27	8681
macro avg	0.51	0.52	0.24	8681
weighted avg	0.91	0.27	0.37	8681

Figure 6. Confusion matrix and classification report of the original data set

Furthermore, we discovered that the Precision of the class 0 (no bankruptcy) and the Recall of the class 1 (bankruptcy) were fairly good based on the indications in the classification report (Figure 6). This demonstrates the model's ability to correctly anticipate bankrupt estimates for the market as a whole. It means the model can find out about t81% companies which will soon be under pressure of financial distress on the market. In addition, the Precision of predictions for non-bankrupt companies is also high, about 96%. So this model is expected to gain the trust of investors, managers and so on. However, the Recall of class 0 and Precision of class 1 are relatively low, indicating that the model overlooks many investment opportunities from companies that have the financial strength in the market but are expected to bankrupt and the model also has a negative look to the market. In conclusion, predicting as many companies going bankrupt as possible was the main aim of this study. As a result of the Classification report, our model has almost achieved that goal.

Formulae for Classification report:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 * precision * recall}{precision + recall}$$

Noticeably, the large difference between Precision and Recall scores of class 1 show that this data set is imbalanced. There is a minor class in bankruptcy and a major class in non-bankruptcy. Therefore, we must resample this dataset to improve this model performance.

4.3.3. Feature importance

	(%)
(CA-Inv)/(S-T lia)	0.068505
sales/receivables	0.053469
[(cash+(S-T securities)+receivables-(S-T lia))/(operating expenses-De)]*365	0.044884
(CA-Inv-receivables)/S-T lia	0.043013
(Inv*365)/COGS	0.041589
operating expenses/total lia	0.040628
(net profit+De)/total lia	0.038751
total lia/((profit on operating activities+De)*(12/365))	0.038173
working capital	0.036538
(gross profit + De)/total lia	0.035856
[CA-Inv-(S-T lia)]/(sales-gross profit-De)	0.035458
(total lia*365)/(gross profit+De)	0.035252
log of TA	0.034588
gross profit/S-T lia	0.033825
CA/S-T lia	0.033735
sales/TA	0.033350
sales/S-T lia	0.032018
constant capital/FA	0.031966
sales/FA	0.031739
book value of E/total lia	0.031376
(S-T lia*365)/COGS)	0.031289
working capital/FA	0.030996
CA/total lia	0.030486
E/FA	0.030228
TA/total lia	0.030166
operating expenses/S-T lia	0.030008
Amount of years until bankrupt	0.021710
L-T Lia/E	0.020405

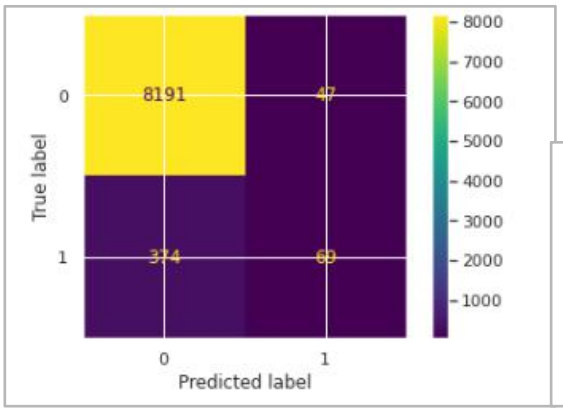
We next examined the model's feature importance. Figure 7 shows that, in general, variables with impact levels range from 2.04% to 6.85%. Furthermore, we can see that short-term debt and highly liquid assets such as cash, receivables, short-term securities, and so on are the primary assets that account for important financial ratios so that it highly affects to the results of operations and financial position of the enterprise. This demonstrates the need of paying more attention to these values, as changes in them in the future might have an impact on the financial strength of businesses.

Figure 7. Feature importance of the original dataset

4.4. Resampling data set

4.4.1. Oversampling

We oversampling the training dataset into a dataset then the numbers of observations between 2 classes are equal which are 33076 observations of each class, and this dataset becomes balanced.



	precision	recall	f1-score	support
0	0.96	0.99	0.97	8238
1	0.59	0.16	0.25	443
accuracy			0.95	8681
macro avg	0.78	0.58	0.61	8681
weighted avg	0.94	0.95	0.94	8681

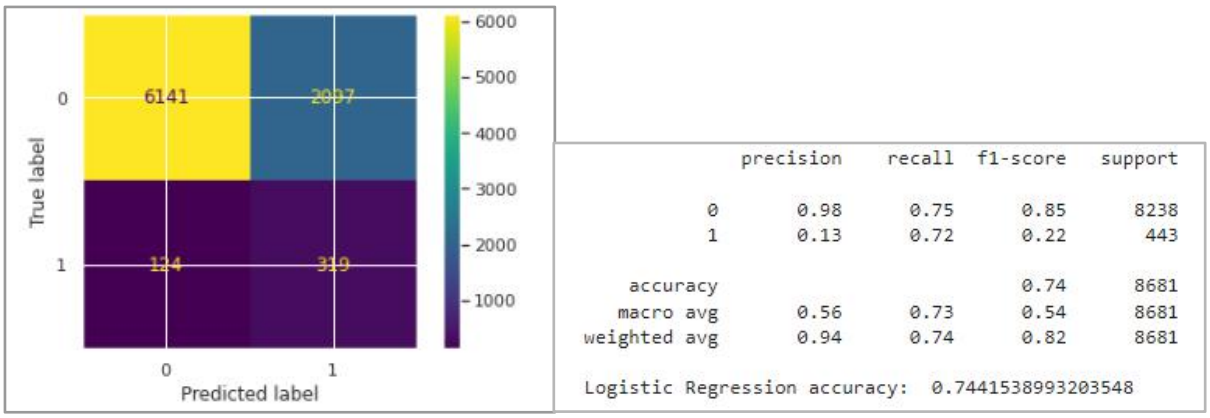
Random Forest accuracy: 0.9515032830319088

Figure 8. Confusion matrix and classification report of the oversampling data set

There is a lack of bankruptcy characteristics in class 1, so the performance of prediction for bankruptcy is very low. In particular, the Precision of class 1 is 59% and the Recall of class 1 is 16%. It means only 16% bankrupt companies on the market are found out by this oversampling model. Therefore we have to do an undersampling method next.

4.4.2. Undersampling

We undersampling the training dataset into a dataset then the numbers of observations between 2 classes are equal which are 1648 observations of each class, and this dataset becomes balanced.



	precision	recall	f1-score	support
0	0.98	0.75	0.85	8238
1	0.13	0.72	0.22	443
accuracy			0.74	8681
macro avg	0.56	0.73	0.54	8681
weighted avg	0.94	0.74	0.82	8681

Logistic Regression accuracy: 0.7441538993203548

Figure 9. Confusion matrix and classification report of the undersampling data set

Following the classification report above, we can see the improvement of the evaluative scores of the model. The Recall of class 0 is up to 75% and the Precision of class 0 increases by 98%. Besides,the total accuracy rises to 74%, insteads of 27% as the original dataset model. However, the main score of class 1 dropped quite a bit. Specifically, the Recall of class 1 decreases from 81% to 72%. It means only 72% of companies which will be bankrupt soon are found out rather than 81% as before. So it is up to the purpose of users and the risky taste of investors to choose the most suitable method.

4.5. Rank the number of years until bankruptcy most correctly predicted

Ranking the number of years to bankruptcy forecast is a useful tool to help users of this model know when is the best time to use financial ratios to predict the company's financial health and whether this company will go bankrupt after that time. Proving this hypothesis will help increase the reliability of the model, save time and cost in the data collection process. Instead of the user having to collect data for up to those 5 subsets, with this forecast time rating tool, users only need to collect data for the fiscal year with the highest rating.

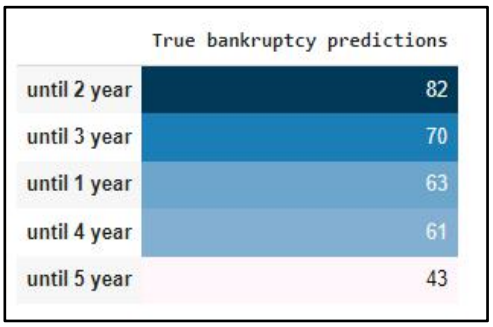
4.5.1. Rank by all true predictions

	All true predictions
until 4 year	1536
until 3 year	1448
until 2 year	1433
until 5 year	1139
until 1 year	904

For ranking the number of years to predict bankruptcy of enterprises based on the total number of correct forecasts on all forecasts, it shows that forecasting 4 years in advance will bring the highest probability of accuracy compared to the 1, 2, 3, or 5 years in advance of forecasting. However, this result is also influenced by the number of observations in each subset because the number of observations in the fourth subset is the largest, and therefore the number of correct predictions is also higher. To overcome this problem, we can use the ratio as a measure of how accurate the rankings are.

Figure 10. Rank the number of years until bankruptcy most correctly predicted by all true predictions

4.5.2. Rank by only true bankruptcy predictions



For ranking the number of years to predict bankruptcy of enterprises based on the total number of true bankruptcy forecasts on all forecasts, it shows that forecasting 2 years in advance will bring the highest probability of accuracy compared to the 1, 3, 4 or 5 years in advance of forecasting. It proved that most companies with financial distress may be under financial pressure in 2 years before bankruptcy. Therefore, users can base on this time rating tool to choose the right financial year for the bankruptcy forecast of the business.

Figure 11. Rank the number of years until bankruptcy most correctly predicted by only true bankrupt predictions

4.6. Change threshold from 50% into 40%

	precision	recall	f1-score	support
0	0.99	0.60	0.75	8238
1	0.10	0.85	0.18	443
accuracy			0.62	8681
macro avg	0.54	0.72	0.47	8681
weighted avg	0.94	0.62	0.72	8681

Figure 12. Classification report of the model with threshold of 40%

With a threshold of 40%, if the observations are predicted to have a bankruptcy probability of less than 40%, they will be classified into class 0 and with a bankruptcy probability of 40% or more, they will be classified into class 1.

After changing the threshold of 40% for this model, we conclude that the threshold significantly improves the model effect to meet the aim. The main intent is predicting as many businesses will actually go bankrupt in the market as possible and the ability to accurately forecast non-bankrupt businesses must be high.

5. Conclusion and discussion

5.1. Conclusion

Results based on a minimum set of independent variables show that our ML method achieves the well-performed of the Precision of the class 0 (no bankruptcy) and the Recall of the class 1 (bankruptcy).

short-term debt and highly liquid assets such as cash, receivables, short-term securities, and so on are the primary assets that account for important financial ratios so that it highly affects to the results of operations and financial position of the enterprise. It is a reasonable and factual result. After resampling dataset, the model performance also becomes better because of healing the imbalanced data set problem. It is true about 95% of non-bankrupt company forecasts and about 72% of bankrupt company in the market found. To make this model more reliable in prediction of the cover of most companies that will go bankrupt in a specific future period for investors to avoid those companies and for managers to change their strategy. However, this model especially works well when making down threshold to 40% rather than 50%, it predicts truely about 99% of non-bankrupt company forecasts and about 85% of bankrupt company in the market found. In addition, if investors are concerned

about the risk of bankruptcy of a business at some point in the future, the bankruptcy prediction after 2 years from the financial year of the collected data is assessed to be best efficiency.

5.2. Discussion

Because of the large number of decision trees involved in the test, random forests are regarded as an accurate and powerful approach. It does not have any overfitting issues. The fundamental reason for this is that it takes the average of all forecasts, canceling out the bias. The approach is applicable to our research. Missing values can also be handled using random forests. We use an effective approach to dealing with these numbers: utilizing linear regression method to replace continuous variables missing values. That may obtain relative feature importance, which aids in determining which characteristics contribute the most to the classifier.

However, this method has some limitations. Since there are so many decision trees in random forests, it takes a long time to make predictions. Every time it makes a forecast, all of the trees in the forest must make a prediction for the same supplied input and then vote on it. This entire procedure takes time. Models are more perplexing than decision trees, which allow you to readily make judgments by following a path through the tree. Additionally, our Machine Learning model was imbalanced and we solved it by oversampling and undersampling methods. We feel that these results were improved which meet our primary goal as many real bankrupt companies which will be predicted as possible. Because the predictive performance achieved on the data set is very good,especially on undersampling dataset.

Possibly this could suggest a connection between financial reporting and the risk of bankruptcy, however, it could also be the case that the distribution of missing values between the successful and unsuccessful companies differs due to the fact that the financial information was collected from two different distributions. Future interesting research topics in the development of algorithms for bankruptcy predictions could be to incorporate new types of data. For example unstructured data, such as texts in annual reports, articles, macro data, and so on could be of interest. For this purpose we can soon apply Vietnam firms data set to make a clear view about the financial health of members in the market.

5.2.1. The applicability of research in management and investment:

Administrators

These outcomes help internal users (such as managers, shareholders and employees). They may identify the drivers of increasing their bankruptcy risk after comparing it to the industry index and general financial index in the results, and so focus more on the elements that improve those ratios to make the performance best. In each specific case of market, companies can asses their and rival's financial health to plan their future strategy. Besides, using time ranking tool nay help user to save their time, their costs for data collection and prove the model reliability.

Investors

Other external users (such as investors, creditors, new established companies, tax authority) also may get advantages from these results. It is clear that those users especially concern about the financial ratios of companies and the determinants of their financial heath to make the best decision.

6. Reference

Esteban Alfaro, Noelia García, Matías Gámez, and David Elizondo. Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems*, 45(1):110–122, 2008.

Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

Edward I Altman. A further empirical investigation of the bankruptcy cost question. *The Journal of Finance*, 39(4):1067–1089, 1984.

Dave Anderson and George McNeill. Artificial neural networks technology. *Kaman Sciences Corporation*, 258(6):1–83, 1992.

Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

William H Beaver. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111, 1966.

Victor M Becerra, Roberto KH Galvão, and Magda Abou-Seada. Neural and wavelet network models for financial distress classification. *Data Mining and Knowledge Discovery*, 11(1):35–55, 2005. Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

Michael P Clements, Philip Hans Franses, and Norman R Swanson. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, 20(2):169–183, 2004.

Ralph De Haas and Neeltje Van Horen. International shock transmission after the lehman brothers collapse: Evidence from syndicated lending. *The American Economic Review*, 102(3):231–237, 2012. A Rogier T Donders, Geert JMG van der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091, 2006.