MASTER 2

ECONOMETRICS, BIG DATA AND STATISTICS (EBDS)

Advanced methods in big data

PREDICTING WHETHER A COMPANY WILL GO BANKRUPT

Authors :

Hadare IDRISSOU

Rachad LIASSOU

Vital GUINGUINNI

Supervized by :

Ewen GALLIC

Sebastien LAURENT

Pierre MICHEL

January 2022

CONTENTS

Introduction

In their development, almost all companies go through debt phases to finance their activities. Some manage to develop, emerge and reach maturity while others have much more difficulty and go bankrupt. Everything starts from a situation of financial distress that translates into an abnormally high debt and can end in total bankruptcy.

The financial institutions that grant loans to businesses try to predict, thanks to the financial variables of the businesses, if they will be able to repay the loans in time. Predicting the failure of a company can therefore be very important for both sides.

On the side of the companies, it would allow to evaluate the internal dynamics and to see where the policies in place lead the company in order to direct new policies, new investments, etc.

For financial institutions, it would allow them to make appropriate lending decisions. The model of predicting the failure of a company was introduced in 1968 by Altman. According to Sun, Li, Huang & He in 2014, mild financial distress can be expressed as a temporary cash flow difficulty, such as the concepts of insolvency, default and the most serious is the bankruptcy of the company.

Therefore, the objective of our work is to use dimensionality reduction techniques and machine learning models to build predictive models of company bankruptcy using its financial information.

Abstract

Our study aims at predicting business failure. It was conducted using data from 6819 companies and 96 variables. After cleaning the database, we ended up with 93 variables. Also, given the imbalance observed in the classes of the variable of interest, we rebalanced the classes in the learning curve before estimating our different models. From our analysis, it appears that high values of the debt ratio and of the current liabilities to assets ratio are reasons that justify the bankruptcy of firms. On the other hand, high or increasing net income to total assets and return on assets allow firms to avoid bankruptcy and perform better. Moreover, the estimation of different models revealed that the random forest is the one that best predicts whether firms will fail or not with an accuracy of 99%, an estimated precision of about 88%, a recall of about 92% and an f1-score of 85%.

I- Literature review

There are several ambiguities about the actual definition of business failure among the various authors who have worked on the subject. Some authors define it simply as the failure of the company, others as financial distress. Thus, for Altman and Hotchikiss in 2006, there are four terms to describe unsuccessful firms: failure, insolvency, default and bankruptcy. They describe failure as the rate of return realized on invested capital that is dramatically and persistently lower than the prevailing rates on equivalent risk-adjusted investments; insolvency as a firm's inability to meet its current obligations; default as a firm's failure to meet an obligation, particularly the payment of a loan or appearance in a lower court; and finally, they believe that there are two types of bankruptcy: one that relates to the net worth of the business and the second that refers to the formal declaration of the business in a federal district court accompanied by a petition either to liquidate its assets or to attempt a reorganization program.

But this definition of bankruptcy by Altman and Hotchikiss does not meet with unanimous approval because for Ross, Westerfield and Jaffe in 1999, there are three types of bankruptcy: legal bankruptcy, technical bankruptcy and accounting bankruptcy. Legal bankruptcy just means that the company goes to court to obtain a declaration of bankruptcy. Technical bankruptcy which describes the situation in which a company cannot fulfill the contract within the stipulated time and accounting bankruptcy which refers to the situation in which a company simply has negative net book assets.

In a somewhat similar sense, Korol & Korodi in 2011, study so-called early warning signals to predict bankruptcy risk.

The different authors had different approaches according to their time and the study techniques developed in their time

Altman was one of the first to introduce a prediction model with a 5-factor discriminant analysis. The main models developed at the beginning were multivariate discriminant analysis, logit and probit models. As statistical techniques improved, new models were developed.

With the rise of artificial intelligence, the neural network became a widely used tool. For example, Pan (2012) wanted to optimize the general regression neural network model by applying the algorithm and obtained good convergence results that indicate the good predictive ability of the model. For example, models such as the support vector machine were introduced in 2005 by Shin, Lee & Kim, 2005; Min & Lee.

1- Materials and methods

1-1- Data description

A company faces bankruptcy when they are unable to pay off their debts. The Taiwan Economic Journal for the years 1999 to 2009 has listed the details of company bankruptcy based on the business regulations of the Taiwan Stock Exchange. The Taiwan Stock Exchange was established in 1961 and began operating as a stock exchange on 9 February 1962. It is a financial institution located in Taipei, Taiwan. It has over 900 listed companies. The data includes a majority of numerical attributes that help understand the possibility of bankruptcy. It contains 6819 companies, 96 attributes, two categories. There is a huge imbalance in the database with 96.774% non-bankrupt firms and 3.226% bankrupt firms. Firms in bankruptcy and not in bankruptcy are marked separately as '1' and '0'. Data are from https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction.

1-2- Appraisal of the database

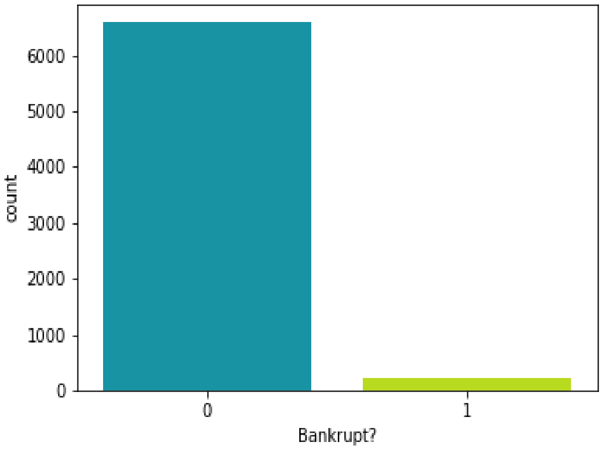
The database initially contains 95 explanatory variables and the dependent variable, which is whether or not the firm is bankrupt. A first exploration of the data through the verification of the missing data showed us that there are no missing values in the database. Then, the exploration of the types of variables revealed that we have 3 categorical variables including the dependent variable and 93 quantitative variables. Through univariate analyses on the qualitative variables of the database, we noticed that apart from the dependent variable, the two others qualitative explanatory variables present very unbalanced classes. All (100%) of the observations are found in a single class for one and more than 99% of the observations are found in a single class for the second variable. We therefore removed these variables from our analysis. Also, since the dependent variable has unbalanced classes, we rebalanced the classes only on the train set within this variable before estimating the different models.

1-3- Descriptive statistics

Univariate analysis

We notice the imbalance between the modalities of the variable of interest, hence the use of a rebalancing technique. In fact, about 3% of the companies in our database have gone bankrupt and 97% are not bankrupt

Figure1: Distribution of firms by business failure



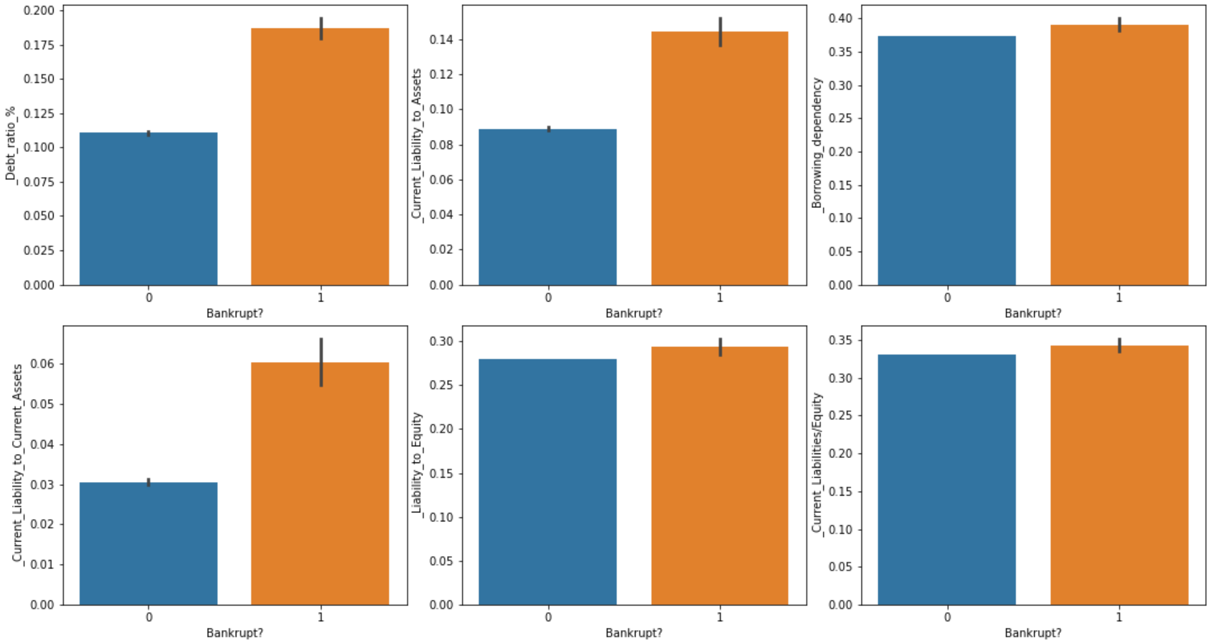
Variable	Outcome	Count	Percent
Bankrupt?	0	6599	96.77
	1	220	3.23

Source : Realized by the authors

Bivariate analysis

We note a strong positive relationship between the features "Debt Ratio_%, Current_Liability_To_Assets, Current_Liability_To_Current Assets" and bankruptcy. Indeed, these indicators are much higher in companies that have gone bankrupt than in those that have not.

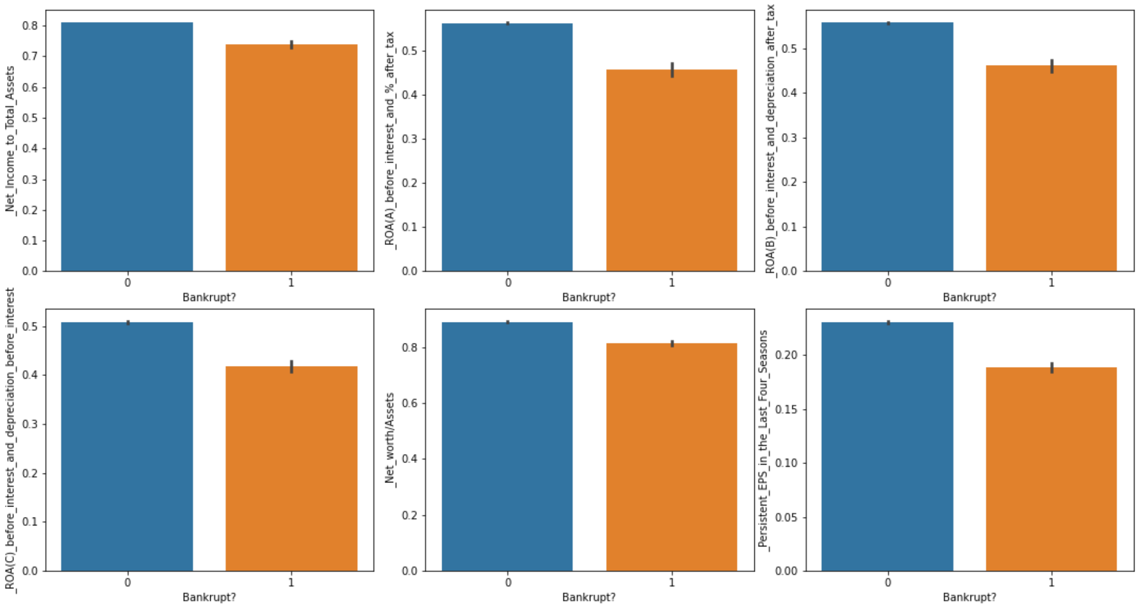
Figure 2: Positive correlation with the target attribute



Source : Realized by the authors

Furthermore, we observe a negative relationship between bankruptcy and features "_ROA(C)_before_interest_and_depreciation_before_interest,_ROA(B)_before_interest_and_depreciation_after_tax, _Net_Income_to_Total_Assets". Indeed, these indicators are lower in firms that have gone bankrupt compared to those that have not.

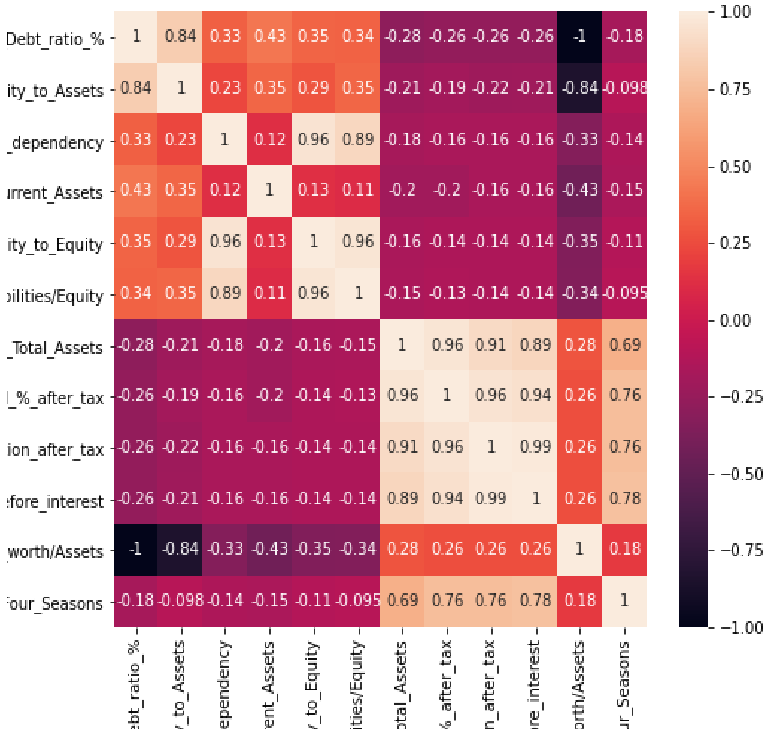
Figure 3: Negative correlation with the target attribute



Source: Realized by the authors

Looking at the correlation matrix, we notice a strong linear relationship between some of our explanatory variables, suggesting problems of multicollinearity.

Figure 4: Correlation matrix



Source : Realized by the authors

1-4- Database rebalancement

As we have 96.774% non-bankrupt firms and 3.226% bankrupt firms, we have to rebalance the database.

In the presence of unbalanced data, resampling is a solution often used to overcome the problem. There are two types of methods: subsampling and oversampling. When there are two unbalanced classes, the preferred operation is oversampling, i.e. generating observations in the underrepresented class until it reaches the second class. This technique is preferred because in the case of under sampling, observations that may contain important information for future analyses are removed. SMOTE is an oversampling technique based on the K-Nearest Neighbors algorithm and allows to generate synthetic data from the existing ones in the dataset. First, the number of observations to be created is defined, an instance is randomly selected and the algorithm is iterated until the classes are rebalanced. At the level of each new observation, values are assigned by multiplying the distance between the point that was used to generate it and it by a value between 0 and 1 and this value is added to the original vector.

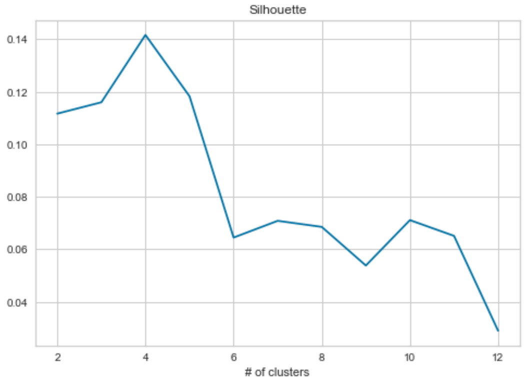
2- Methods and results

2-1- Unsupervised method: K-Means

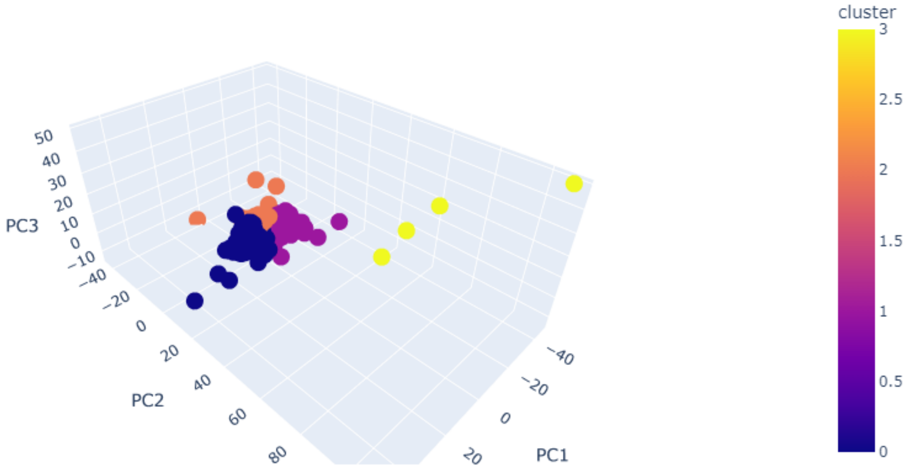
Unsupervised clustering algorithm, the K_means attempts to divide into K distinct clusters or K groups the observations according to their similarities. The principle consists in first choosing randomly K centroid, which are the centers of the clusters. Then, each point is assigned to the group, which is closest to the center in terms of distance, and then we recalculate the center of each cluster and modify the centroid. The process is repeated until there are no more changes in the centroid values.

In the case of our study, we will use this method to check how bankrupt and non-bankrupt firms behave in the different clusters and compare each category of firm in the clusters.

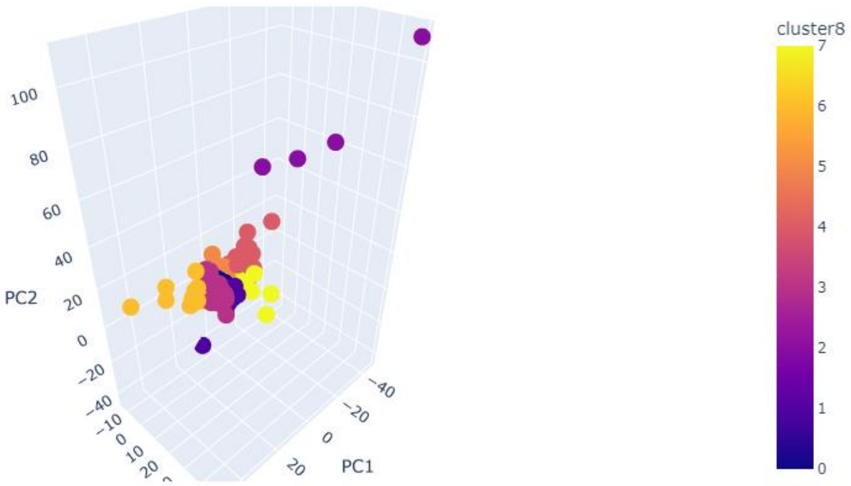
In order to find the optimal number of clusters, we used the silhouette metric, which measures the performance of the clusters. To do so, we varied the number of clusters from 2 to 12. The result is as follows



It is clear that the optimal number of clusters is equal to 4. Prior to the graphical representation and visualization of the clusters, we proceeded to the dimensionality reduction by using the principal component analysis and retained three dimensions. Thus, the graphical representation of the different clusters is as follows.



The visualization through this graph shows that the observations are mostly well grouped according to the optimal cluster number.In order to confirm this, let's try with the default settings of scikit_learn which uses a number of cluster equal to 8 by default.



We note that the clusters appear less separated with 8 clusters and that some groups can be merged very well.

We will now check some characteristics of the clusters according to the dependent variable

Bankrupty	N° clusters	Debt ratio mean	Net_income_to_total_assets mean
0	**0**	0.091580	0.843177
	1	0.148912	0.782147
	2	0.101731	0.807213
	3	0.292711	0.761793
1	**0**	0.156241	0.819091
	1	0.192823	0.721550
	2	0.166321	0.787733
	3	0.293830	0.621947

We find that in all four clusters, the average debt ratio is higher for firms that have failed than for those that have not. On the other hand, the return on total assets, which shows how efficiently a firm uses its assets to generate profits, is on average smaller for bankrupt firms than for non-bankrupt firms.

2-2- Logistic regression

The explanatory analysis will be done using multivariate logistic regression, which allows us to estimate the occurrence of an event by taking into account which it is possible to estimate the occurrence of an event by taking into account auxiliary auxiliary information. The choice of statistical model is determined by the nature of the data to be analyzed their structure and the objectives assigned to our study. Moreover, this model based on the probability estimate P(Y=1/X) by maximum likelihood has the advantage of being the most suitable for qualitative studies among all the models on binary qualitative variables (Bourbonnais, 2009). The explained variable, also called dependent variable, is dichotomous (firms that have gone bankrupt or not). The model seeks to establish a functional relationship of the form: $Y=$ (X), where Y is the variable to be predicted and Xi, (i =1,2,3, 4...n) the predictor variables. As stated, The nature of the variable Y is dichotomous.

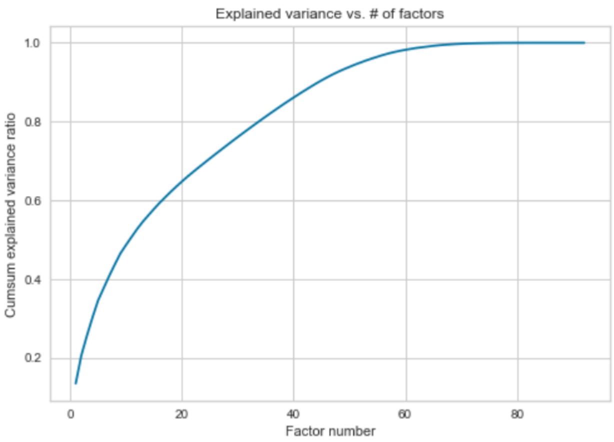
$$Y = \begin{cases} 1, if\ the\ company\ went\ bankrupt \\ 0, sinon \end{cases}$$

Given the very large number of variables at our disposal and the diagnosis of the multicollinearity problem through the multicollinearity which, through the correlation matrix, has revealed its presence, we are considering information reduction techniques that will try to overcome this problem. To do so, we compare two techniques of dimension reduction, namely PCA, LASSO, RIDGE and ELASTIC NET.

2-1-1- Logistic regression based on PCA

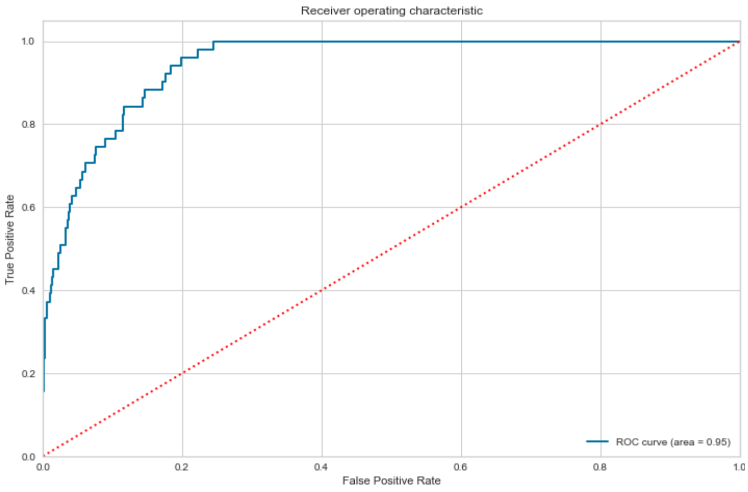
Principal Component Analysis is an exploratory statistical tool. It allows us to explore multi-variate data, that is, data with several variables. Therefore, it is considered as a multivariate statistical analysis or a multivariate technique, allowing to reduce a set of initial variables into a few factors or principal components which are new variables. The original variables are thus reduced into a small number of new variables or principal components, while retaining a maximum of information. There are thus a small number of factors that explain most of the variance contained in the original variables. The information contained in the original variables is thus extracted, visualized and synthesized into a few new variables, from a cross between several numerical variables. The dimensions of a multi-variate data are thus reduced to a few principal components that can be visualized graphically. PCA results in a graphical representation of the data (scatter plot) in relation to these factors or principal components in the form of axes. These axes or principal components summarize as best as possible the set of initial variables. They are linear combinations of the initial variables, hierarchical and independent of each other. To avoid differences in units, the variables must be standardized.

Based on the cumulative sum variance ratio criterion, we took the first 60 factorial axes that contain more than 98% of information.



To evaluate our logistic regression based on PCA, we use the confusion matrix to highlight criteria such as accuracy, precision, recall and also the ROC curve

	Predicted	
Actual	Negative	Positive
Negative	1165	148
Positive	11	40



The accuracy of the model is about 88%. This implies that the model, which seems satisfactory, very well predicted 88% of the companies. However, the precision of the model is about 21%, which leaves something to be desired. Digging a little deeper, we observe that the recall is about 78%, which is not bad since it implies that the model predicted very well or correctly classified 78% of the firms that actually failed. Further on, we also note that although the area under the ROC curve is 95%, the f1-score associated with the model is about 33%. This is not satisfactory.

2-1-2- Logistic regression on Lasso

Lasso stands for Least Absolute Shrinkage and Selection Operator. It reduces the regression coefficients toward zero by penalizing the regression model with a penalty term called the L1 norm, which is the sum of the absolute coefficients.

In the case of lasso regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly zero. This means that the lasso can also be considered as an alternative to subset selection methods for performing variable selection to reduce model complexity.

	Predicted	
Actual	Negative	Positive
Negative	1134	179
Positive	6	45

The accuracy of the model is about 86%. This implies that the model, which seems satisfactory, very well predicted 86% of the companies. However, the precision of the model is about 20%, which leaves something to be desired. Digging a little deeper, we observe that the recall is about 88%, which is not bad since it implies that the model predicted very well or correctly classified 88% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 33%. This is not satisfactory.

2-1-3- Logistic regression on Ridge

Ridge regression reduces the regression coefficients, so that variables with a minor contribution to the outcome have their coefficients close to zero.

The narrowing of the coefficients is achieved by penalizing the regression model with a penalty term called the L2 norm, which is the sum of the squared coefficients. The amount of the penalty can be refined using a constant called lambda (λ). Choosing a good value for λ is critical. As λ increases to infinity, the impact of the shrinkage penalty increases and the peak regression coefficients approach zero.

Ridge regression reduces the coefficients toward zero, but it will not set any of them exactly to zero.

	Predicted	
Actual	Negative	Positive
Negative	1171	142
Positive	8	43

The accuracy of the model is about 89%. This implies that 89% of the companies were very well predicted by the model, which seems satisfactory. However, the precision of the model is about 23%, which leaves something to be desired. Digging a little deeper, we observe that the recall is about 84%, which is not bad since it implies that the model predicted very well or correctly classified 84% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 36%. This is not satisfactory.

2-1-4- Logistic regression on Elasticnet

Elastic Net produces a regression model that is penalized by both the L1 and L2 norms. The consequence of this is to effectively reduce the coefficients (as in RIDGE regression) and set some coefficients to zero (as in LASSO).

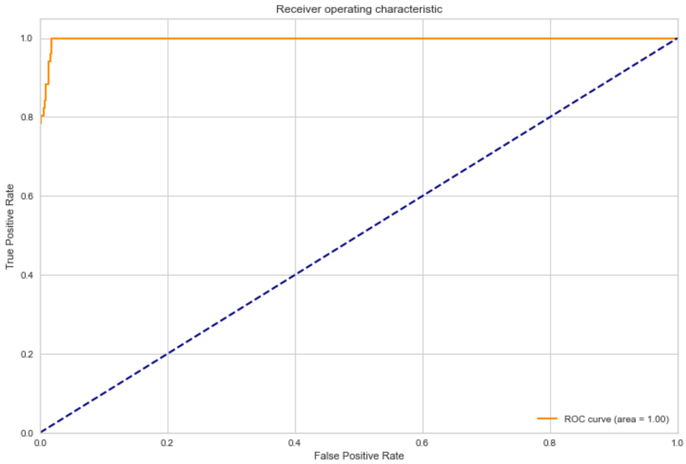
	Predicted	
Actual	Negative	Positive
Negative	1171	142
Positive	8	43

The accuracy of the model is about 89%. This implies that 89% of the companies were very well predicted by the model, which seems satisfactory. However, the precision of the model is about 23%, which leaves something to be desired. Digging a little deeper, we observe that the recall is about 82%, which is not bad since it implies that the model predicted very well or correctly classified 82% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 36%. This is not also satisfactory.

2-3- Support Vector Machine (SVM)

SVM belong to a family of algorithms that use so-called supervised learning, and are specialized in solving mathematical problems of discrimination and regression. SVMs are considered a generalization of linear classifiers. In SVMs, data are separated into multiple classes using the "maximum margin", with a separation boundary chosen to maximize the distance between groups of data. SVM (Vector Machine Support) is a supervised machine learning algorithm that can be used for both regression and classification. The idea of this method is to find the hyperplane that optimally separates the data while maximizing the margin. In other words, a hyperplane is the one that maximizes, in the case of this study, the distance between the group of failed firms and the group of non-failed firms. The hyperplane is referred to as the line that correctly classifies the data. The vector supports are the values in the data set that are close to the hyperplane.

	Predicted	
Actual	Négatif	Positif
Négatif	1307	6
Positif	10	41



The accuracy of the model is about 99%. This implies that the model, which seems satisfactory, very well predicted 99% of the companies. Also, the precision of the model is about 87%, which is satisfactory. Digging a little deeper, we observe that the recall is about 80%, which is not bad since it implies that the model predicted very well or correctly classified 80% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 84%. This model seems very satisfactory in view of its performance.

2-4- K-Nearest Neighbors

The KNN algorithm belongs to the class of supervised learning algorithms and is used to solve classification and regression problems. It aims to classify target points according to their distances from points constituting a training sample, the number K represents Neighbors.

From a randomly chosen point, circles are created around it like a radar to scan and find close neighbors belonging to the same category.The hyperparameter k allows to find the number of neighbors that must be inside the circles whose center is the chosen point.

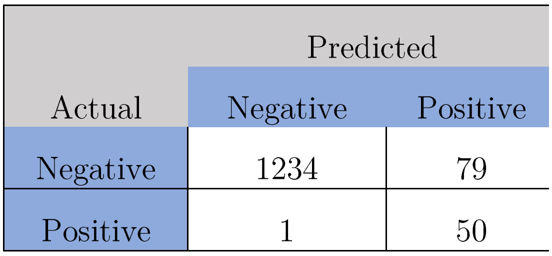
To determine the optimal hyperparameters, we initiate a process through the GridSearchCV that allows us to determine the optimal parameters :

N_neighbors : The number of neighboring points that must be inside the circle.

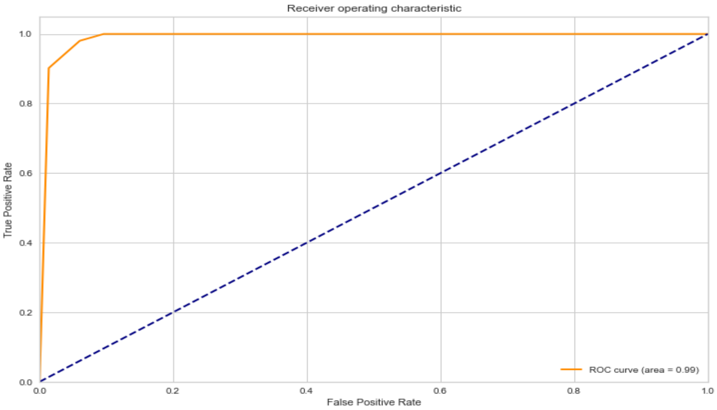
Weights : which can be set to either 'uniform', where each neighbor within the boundary carries the same weight or 'distance' where closer points will be more heavily weighted toward the decision.

Metric : which refers to how the distance of neighboring points is chosen from the unknown point.

The Grid Search give us the following hyper parameters : neighbors : 4, weights : uniform, metric : Manhattan.



	Predicted	
Actual	Negative	Positive
Negative	1234	79
Positive	1	50



The accuracy of the model is about 94%. This implies that the model, which seems satisfactory, very well predicted 94% of the companies. Also, the precision of the model is about 39%, which is not satisfactory. Digging a little deeper, we observe that the recall is about 98%, which is not bad since it implies that the model predicted very well or correctly classified 98% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 56%. This model seems satisfactory in view of its performance.

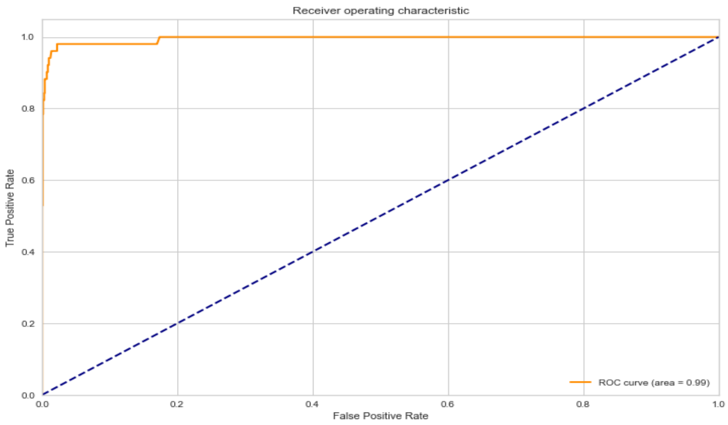
2-5- Random Forest Classifier

The random forest or decision tree forest is a particularly efficient algorithm in terms of predictions in the field of machine learning. It is composed of several decision trees that each produce an estimate whose assembly allows to obtain a global estimate. It is based on the bagging system and the final estimate is made from a classification method that consists in choosing the category that comes up most often.

To apply the random forest, we have to choose the number of decision trees and the number of variables to implement. A Grid search is used to test a series of parameters and to identify the most useful ones.

The Grid Search allowed us to select 26 trees and 48 variables for our estimation:

	Predicted	
Actual	Negative	Positive
Negative	1301	12
Positive	4	47



The accuracy of the model is about 99%. This implies that the model, which seems satisfactory, very well predicted 99% of the companies. Also, the precision of the model is about 80%, which is satisfactory. Digging a little deeper, we observe that the recall is about 92%, which is not bad since it implies that the model predicted very well or correctly classified 92% of the firms that actually failed. Further on, we also note that the f1-score associated with the model is about 85%. This model seems very satisfactory in view of its performance.

3- Comments

From this table, we can see that the parametric methods, although offering advantages in terms of interpretation, present precision scores (between 20% and 23%) and f1-score (between 33% and 36%) that leave something to be desired. In the framework of this study, they provide performances well below the non-parametric models.

As far as the parametric methods are concerned, they present satisfactory results in all score points even if the precision (39%) of the K-nearest neigbors leaves something to be desired. Two models clearly stand out by their performances. They are the support vector machine with a precision of 87%, a recall of 80% and an f1-score of 84% and the random forest with an precision of 80%, a recall of 92% and an f1-score of 85%.

In order to distinguish between the two models, as they have the same score for predictive power and as one model outperforms the other in terms of accuracy and vice versa in terms of recall, we used the f1-score. In this game, the random forest appears in the lead with a score of 85% against 84% for the vector machine support. We thus retain this method as being the best.

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
LOGISTIC REGRESSION ON PCA	0.88	0.21	0.78	0.33
LOGISTIC REGRESSION ON LASSO	0.86	0.20	0.88	0.33
LOGISTIC REGRESSION ON RIDGE	0.89	0.23	0.84	0.36
LOGISTIC REGRESSION ON ELASTICNET	0.89	0.23	0.82	0.36
SUPPORT VECTOR MACHINE	0.99*	0.87*	0.80	0.84
K_NEAREST NEIGHBORS ALGORITHM	0.94	0.39	0.98*	0.56
RANDOM FOREST ALGORITHM	0.99*	0.80	0.92	0.85*

Although Random Forest was chosen as the best model, depending on the objectives of a potential user, we could choose SVM because of its performance in terms of accuracy.

Conclusion

The biggest disappointment a company can experience is bankruptcy. The present study, which aims to predict business failure, is based on business data collected by the Taiwan Economic Newspaper. A total of 6819 companies are considered for this study. From the results of this study, it can be seen that high values of debt ratio or loan dependency are indicators of poor business health and lead to business failure. At the same time, a high level of return on assets is a factor that stimulates the growth of a firm. Furthermore, through this study,we estimated parametric logistic regression models based on PCA, LASSO, RIDGE and ELASTIC NET to reduce the dimensionality of our dataset, and then non-parametric methods such as support vector machine, k-nearest neigbors and random forest to predict the bankruptcy or not of firms in several aspects and to be able to choose the optimal model.

The results of the analyses show that non-parametric models perform much better than parametric models in predicting whether a firm will fail or not. Two models clearly stand out for their very satisfactory performance. They are the support vector machine with a precision of 87%, a recall of 80% and an f1-score of 84% and the random forest with an precision of 80%, a recall of 92% and an f1-score of 85%. Finally, the random forest was selected as the best model.

II- Bibliography

1- Abbas S, Jalil Z, Javed AR, Batool I, Khan MZ, Noorwali A, et al. BCD-WERT: A Novel Approach for Breast Cancer Detection Using Whale Optimization Based Efficient Features and Extremely Randomized Tree Algorithm. PeerJ Computer Science. 2021;7:e390.

2- Sun J, Li H, Huang QH, He KY. Predicting Financial Distress and Corporate Failure: A Review from the State-of-the-Art Definitions, Modeling, Sampling, and Featuring Approaches. Knowledge-Based Systems

3- Ohlson JA. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research.

4- Altman EI. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance.

5- Beaver WH. Financial Ratios As Predictors of Failure. Journal of Accounting Research.

6- Shi Y, Li X. An Overview of Bankruptcy Prediction Models for Corporate Firms: A Systematic Literature Review. Intangible Capital.

7- Lahmiri S, Bekiros S. Can Machine Learning Approaches Predict Corporate Bankruptcy? Evidence from a Qualitative Experimental Design. Quantitative Finance.

8- Singh A, Purohit A. A Survey on Methods for Solving Data Imbalance Problem for Classification. International Journal of Computer Applications

9- Edward I. Altman, Edith Hotchkiss ;Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy, Analyze and Invest in Distressed Debt, Third Edition

III- Appendix

- The confusion matrix

The confusion matrix is a matrix that measures the quality of a classification model. The rows correspond to the real classes and the columns to the predicted classes. Thus, on the diagonal, the correct predictions for the different classes are represented.

	Predicted	
Actual	Negative	Positive
Negative	TN	FP
Positive	FN	TP

The accuracy measures the rate of good predictions on the test sample :

$$Accuracy = \frac{TP\ +\ TN}{TP\ +\ FP\ +\ TN\ +\ FN}$$

The accuracy gives the rate of good predictions for observations predicted to be in the class of interest.

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

The recall gives the rate of good predictions for the observations of the class of interest

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

The F1-Score is a compromise between recall and precision

$$F1\ Score = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

- <u>Database</u>