

Sustainable and Responsible Management

The Journal of Sustainable Business



Volume No. 2, Issue No. 1 - 2021

Abruzzese Salute Mutua

sustainable-and-responsible-management.org

Journal Information

Editor-in-Chief

Michele Samuele Borgia

Managing and Responsible Editor

Maurizio Cirillo

Copy Editors

Stefano Cesinaro

Katia Di Tommaso

ISSN

2724-4466

Online since 18th November, 2020

With authorization from the Court of Pescara no. 1 year 2020

Copyright Holder

Abruzzese Salute Mutua

Via Umberto I n. 78/80

65010 Cappelle sul Tavo (PE)

Italy

Statistical Models for Probability of Default: An Empirical Application to Italian Food Industry

Carlo Zaccardi *

Department of Philosophical, Pedagogical and Economic-Quantitative Sciences
University “G. d’Annunzio” of Chieti-Pescara
E-mail: carlo.zaccardi@unich.it

** Corresponding author*

Abstract

This paper discusses three different statistical approaches to estimate the *probability of default (PD)*, which represents the likelihood of a company to become insolvent within a certain future time, and is an essential component when evaluating the credit risk backed by a credit institute: due to the presence of information asymmetry, when a bank lends money, it is likely that the debtor will not pay back its past due debt; one of the reasons of such a possibility is that the company-borrower becomes insolvent. The three approaches are the linear discriminant analysis (LDA), the logistic regression, and the classification tree: each model is applied using empirical data from Italian food industry. The statistical methods discussed in the paper belong to the econometric models’ family, since they attribute a *score* to each firm: thanks to a dimensionality reduction process, balance sheet indexes are transformed into a value, which is used by the bank in order to know if the firm-client has the right characteristics to receive the loan. The value corresponds to the probability of default in the case of logistic regression and classification tree models. Empirical analyses are conducted using the software *R* and show that the statistical models lead to quite similar results, although they have many differences from the theoretical point of view. Eventually, a comparison between the models shows that the classification tree is the best one. But is it always better than the other two approaches? The answer is negative.

Keywords – Credit risk; Probability of default; R; Linear discriminant analysis; Logistic regression; Classification tree.

Paper type – Academic Research Paper

Article history:

Received 16 January 2021

Received in revised form 15 March 2021

Accepted 30 April 2020

Available online

1 Introduction

Banks are exposed to a variety of financial risks, but the most important is the credit risk, since loans are banks' principal product. This risk provoked several economic disasters in the last decades: one of the main reasons of these recessions is that credit institutes held a too large fraction of non-performing loans and they had underestimated the importance of the relative risk (Naili & Lahrichi, 2020). So, correctly assessing the real importance of credit risk is fundamental, not only for the single bank, but also for the entire economic and financial system. In fact, credit institutes are interconnected, thus a single collapse could produce a "domino effect" whose entity might be so important that the global financial system will fall apart (Badarau & Lapteacru, 2020; Naili & Lahrichi, 2020).

This paper aims to analyse, from a statistical perspective, how the management of credit institutes can assess if a client (i.e. a company that requests a loan) is reliable and how much it is reliable.

It is assumed that the legal representative of the company set his or her foot in the credit institute for the first time. The presence of information asymmetry is the main problem in this case: the creditor (the bank) does not know a priori all the information that the debtor (the client) has and does not know if the debtor will act in line with the creditor's interest, so evaluating the firm's reliability becomes a task harder than the case when there is a long term relationship between bank and client (Puri et al., 2011).

However, assessing the trustworthiness of the client depends on the reliability of the model. The perfect model does not exist, so another concern arises: is this reliability high enough to correctly assess that of the client? There are some tests to be run to answer this question (Coolen-Maturi & Coolen, 2019; Witzany, 2017).

In statistical jargon, the probability of default (PD) is the key argument of this paper: it explains how likely the counterparty will become insolvent within a certain future date; moreover, it is a crucial part of the credit risk's evaluation process, alongside with the "loss given default" and the "exposure at default" (Zamore et al., 2018), which will not be treated in this article.

To start with, Zamore et al. (2018) and Witzany (2017) illustrate various models to assess this risk: in particular, Zamore et al. (2018) distinguish between two "families", i.e. qualitative and quantitative models. The first group focuses on formulating a subjective judgement, while the second one aims at producing a score, i.e. a synthetic evaluation of the creditworthiness, useful to estimate the PD. Zamore et al. (2018) and Witzany (2017) briefly describe different quantitative models: econometric models, neural networks, optimization models, expert systems, hybrid systems, structural and reduced-form models, shadow rating models, linear programming and Support Vector Machines. In particular, the approaches that will be discussed in this paper (linear discriminant analysis, logistic regression and classification tree) are included amongst the econometric techniques (Witzany, 2017). They "estimate

the probability of default as the target (explained) variable by a set of explanatory variables including financial ratios, categorical variables, and also qualitative assessment indicators. The models' parameters are estimated based on historical data with information on defaults and the explanatory variables values" (Witzany, 2017, p. 37). In the specific case of this article, only some economic and financial indicators will be taken into account.

To understand the definition of the term "default" adopted in this paper, it is necessary to say what is established by the Article 5 of the Italian Bankruptcy Law (Regio decreto 16 marzo 1942, n. 267). According to it, the insolvency arises through non-fulfilments and other exterior facts which prove that the borrower can't regularly satisfy its obligations, but one single non-fulfilment does not imply being defaulted. The concept of insolvency, as stated by the Article 1 of the same Law, is not applied to those businesses that satisfy simultaneously certain requirements.

However, this definition of default is too narrow for the purpose of this article: it would cost the exclusion of many small firms and the results would become distorted. Because of this problem and the lack of information, when it is not clear whether the default has effectively happened, it is presumed that it happened; so, the definition of "default" adopted includes the following situations: insolvency in the strict sense, bankruptcy, liquidation (not voluntary), controlled administration, total termination, debt reorganization agreement, composition with creditors, and dissolution.

2 Some theoretical aspects

Three methods are discussed in this article: the linear discriminant analysis (LDA), the logistic regression, and the classification tree. They are all classification methods, since the response variable is categorical; the aim is to distinguish amongst units belonging to two different groups, i.e. there are two labels: "defaulted" and "non-defaulted" businesses (Witzany, 2017; James et al., 2013).

Why these three approaches? LDA is a frequently used method (Witzany, 2017; Adnan Aziz & Dar Humayon, 2006), alongside with the logistic regression, which is the most widely used model in banking practice (Witzany, 2017). Then, decision trees¹ are becoming increasingly popular (Jiao et al., 2020); moreover, they are easy to explain and to interpret, and they are considered to be close to the human's decision-making process (James et al., 2013).

¹ There are two categories of decision trees: the classification trees are used when the response variable is categorical, meanwhile the regression trees are used for the other circumstances (James et al., 2013).

LDA (Fisher, 1936) was used for the first time in bankruptcy prediction of companies by Altman (1968). It seeks for a linear function such that the distance between these groups is maximum (Witzany, 2017); consequently, a threshold is calculated, which allows to predict what label an out-of-sample firm has, and to compute the probability of default (Resti & Sironi, 2008). Thus, in this case the dimensionality reduction process transforms balance sheet indexes into a “score” that does not represent the PD, but a further formula should be applied to obtain the probabilities (Resti & Sironi, 2008). But in order to apply this transformation, LDA lies on three assumptions (James et al., 2013):

- Input data, conditionally to each group, distributes normally.
- The variance-covariance matrix is the same between the groups.
- The variables (i.e. balance sheet indexes) are independent from each other.

These assumptions can make LDA difficult to apply (Eisenbeis, 1977), and Zavgren (1985) warns against using this method when the assumptions are violated (Mihalovič, 2016). Nevertheless, Kim et al. (2011) claim the benefits of LDA, while Hastie et al. (2009), based on their experience, state that this method and the logistic regression give similar results, even if the assumptions are not matched.

The logistic regression model (Cox, 1958) is utilized in many fields, e.g. epidemiology, medicine, modelling income, and also in finance (Whitaker et al., 2019). Ohlson (1980) was the first who applied this approach to predict the default event (Mihalovič, 2016).

In comparison with linear regression, logistic regression has better properties, since it estimates an exponential function that gives back values within the [0,1] interval, so that they can be interpreted as probabilities of default (James et al., 2013; Witzany, 2017). The maximum likelihood method is used to fit both LDA and logistic models (Hastie et al., 2009).

The classification tree looks for some regressors that best separate the observations into the two groups, using a specific algorithm. The probabilities of default are carried out for each node of the tree (James et al., 2013).

LDA and the logistic regression work in the following way: they compute the scores at first, and then they allow the creation of the two groups, according to a classification rule. Instead, the classification tree produces the scores (actually, they coincide with the PDs) after having gathered the units (Giudici & Figini, 2005).

Hunt (1962) proposed some basic concepts of the decision trees, but the first regression tree model was proposed by Morgan & Sonquist (1963), then Messenger & Mandell (1972) extended the ideas of Morgan & Sonquist to classification. However, initially these proposals had not been seen as interesting (Loh, 2014). Only a decade later, Breiman et al. (1984) regenerated the interest in the subject thanks to their CART (Classification And Regression Trees) algorithm: for the first time, continuous variables are used to build a tree (Jiao et al., 2020). There are many other algorithms illustrated in Loh (2014), such

as CHAID (Kass, 1980), FACT (Loh & Vanichsetakul, 1988), C4.5 (Quinlan, 1993), QUEST (Loh & Shih, 1997), CRUISE (Kim & Loh, 2001), CTREE (Hothorn et al., 2006), and GUIDE (Loh, 2009).

This paper focuses on the CART algorithm because it is the most popular among statisticians (Alharan et al., 2017; Giudici & Figini, 2005; Wu et al., 2008): the predictor space is divided into several regions using a recursive approach; the algorithm grows a large tree, then the tree is pruned at some point where a cost function is minimised (Loh, 2014).

3 The linear discriminant analysis (LDA)

3.1 Preparing the dataset

As basic notation, let Y be the variable, representing the default event, with two levels: it is equal to 0 if the default has not happened yet, 1 otherwise. Let G_0 be the group of “good” (or “healthy” or “non-defaulted”) companies, and G_1 the group of “bad” (or “unhealthy” or “defaulted”) businesses.

The dataset, updated to 23rd March 2020, is extracted from “Aida”, a data warehouse of the society Bureau van Dijk², and contains all the companies of the Italian food industry, whose output is every kind of food for humans and animals. According to the European classification of economic activities (NACE Rev. 2), these firms belong to the section C, division 10, named “Manufacture of food products” (European Communities, 2008).

Data have been collected with reference to the year 2018, while information about the insolvency is referred to the year 2020; therefore, the purpose of the analysis is to assess the probability of default in 2022. Obviously, PDs may result higher due to the new global crisis unleashed by the virus SARS-CoV-2.

In the first place, the number of balance sheet ratios contained in the dataset have been shrunk picking those which have the least number of empty cells inside the dataset. Here is the complete list with their definitions given by “Aida”:

- *Fixed assets coverage ratio:* $\frac{\text{Total fixed assets}}{\text{Equity}}$

It indicates the equilibrium between fixed and net assets; fixed assets are normally funded by equity, so this ratio should be approximately equal to 1 (Cerved, n.d.);

² <https://www.bvdinfo.com/en-gb>

- **Quick ratio:** $\frac{\text{Current assets} - \text{Inventory}}{\text{Current liabilities}}$

It shows the solvency capacity of a company; a value closer to 0 rather than 1 indicates that the business is potentially unable to pay its debts back in the short-run (Cerved, n.d.).
- **Return on assets (ROA):** $\frac{\text{EBIT}}{\text{Total assets}}$

It represents the yield rate of the equity; EBIT is for Earnings Before Interests and Taxes (Cerved, n.d.);
- **Indebtedness ratio:** $\frac{\text{Total assets}}{\text{Equity}}$

It indicates the amount of assets funded by equity; a higher value represents a higher indebtedness level;
- **Short-term indebtedness ratio:** $\frac{\text{Current debts}}{\text{Total debts}}$

It shows the proportion of debts that the firm has to pay back in the short-run;
- **Bank debts on current assets:** $\frac{\text{Bank debts}}{\text{Current assets}}$

It represents the degree of current assets hedged through bank loans; a value close to 1 indicates the optimal situation, otherwise, if it is smaller than 1, a part of current assets is funded using the equity, while if it is greater than 1, the firm supports its fixed assets using short-run debts (Cerved, n.d.);
- **Financial charges on EBITDA:** $\frac{\text{Financial charges}}{\text{EBITDA}}$

It corresponds to the degree of absorption of economic resources (generated by the company's core business) due to financial charges; a value greater than 1 represents a bad health status (Cerved, n.d.);
- **Assets turnover ratio:** $\frac{\text{Turnover}}{\text{Total assets}}$

It expresses the mean return per unit of investment (Cerved, n.d.);
- **Current assets turnover ratio:** $\frac{\text{Turnover}}{\text{Current assets}}$

It shows the mean return per unit of current assets (Cerved, n.d.);

- Rigidity ratio: $1 - \frac{\text{Current assets}}{\text{Total assets}} = \frac{\text{Total fixed assets}}{\text{Total assets}}$

It indicates the allocation of total assets between fixed and current assets; a value closer to 1 depicts high rigidity, since the company must hoard a lot of money for amortisations (Biggeri et al., 2012);

- Return on debts (ROD): $\frac{\text{Financial charges}}{\text{Total debts}}$

It corresponds to the mean interest rate on active debts (Biggeri et al., 2012);

- BETA: $\frac{\text{Total debts}}{\text{Equity}}$

It compares two types of liabilities: third party and own resources (Biggeri et al., 2012).

Then, the definition of default has been applied to separate the enterprises: as a result, the group of defaulters ($Y=1$) has 549 observations, while there are 11419 non-defaulters ($Y=0$) in the dataset, so extracting a sample has been necessary.

From the entire population, a stratified sample has been extracted using regions and dimension classes as strata. Region classes correspond to level 1 of NUTS (Nomenclature des Unités Territoriales Statistiques) classification, while dimension classes (micro, small, medium-sized and big) have been made from the “Small and Medium-sized Enterprise” (SME) definition adopted by the European Commission (n.d.).

The stratified sample that has been extracted among the good borrowers is 10% of them. Then, the entire sample (containing both defaulters and non-defaulters) has been split between training and test sets; at the same time, data has been standardised.

3.2 Selecting the independent variables

It is necessary to seek for the variables with the highest discriminative capacity. The backward elimination is the procedure that has been chosen, so the aim is to remove the variables with the highest p-value for the F-to-remove test (Resti & Sironi, 2008).

In R, these three functions have been written: they are needed to compute, in order, the Wilk’s Lambda (Resti & Sironi, 2008), the F-to-remove test, and to suggest which variables should be removed.

The latter function is the most important: it tries to identify, one by time, which is the independent variable that produces the highest p-value (or the lowest value of the statistic) when running the F-to-remove test.

After some steps, the tests on the following remaining eight variables say that they should not be removed because they are statistically significant, given a confidence level of 5%.

Table 1 - p-values from F-to-remove test

Variable	p-value
Bank debts on current assets	0.03142
ROD	0.00981
Indebtedness ratio	0.00698
Quick ratio	2.24375e-05
Current asset turnover ratio	2.67355e-07
Asset turnover ratio	1.48241e-10
ROA	7.80424e-12
Fixed assets coverage ratio	4.82711e-20

Source: our elaboration

3.3 Fitting the model

Now the linear discriminant analysis can be run, and in Table 2 are stored the estimates for each input variable.

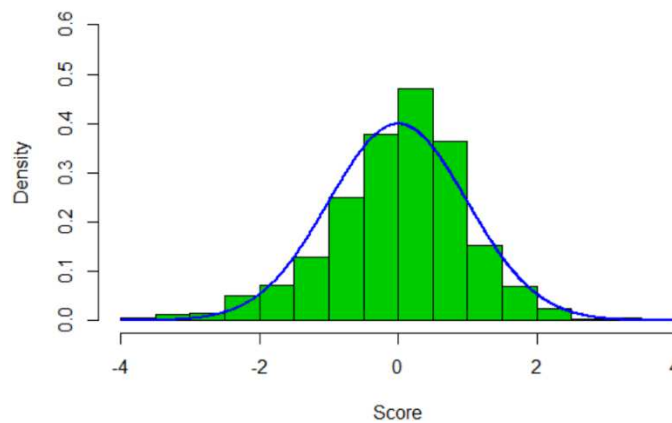
Table 2 - Estimated coefficients from fitted LDA model

Variables	Estimates
ROA	-0.4157480
Quick ratio	-0.3001004
Fixed assets coverage ratio	-0.3518421
Indebtedness ratio	0.4481402
Asset turnover ratio	-0.5155876
Current assets turnover ratio	-0.4316197
Bank debts on current ratio	0.1094183
ROD	-1.2106993

Source: our elaboration

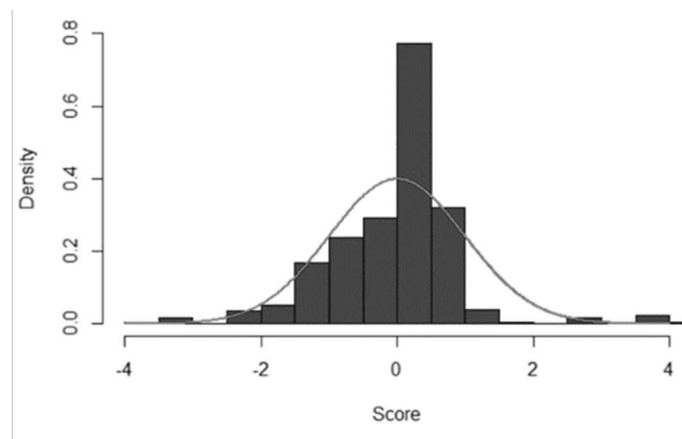
Frequency histograms are plotted and then a line indicating the theoretical standard normal distribution is added. Figure 1 and Figure 2 show that there are both skewness and heavy tails in the case of bad companies, instead the other distribution is quite normal.

Figure 1 - Frequency histogram of the scores obtained by healthy companies



Source: our elaboration

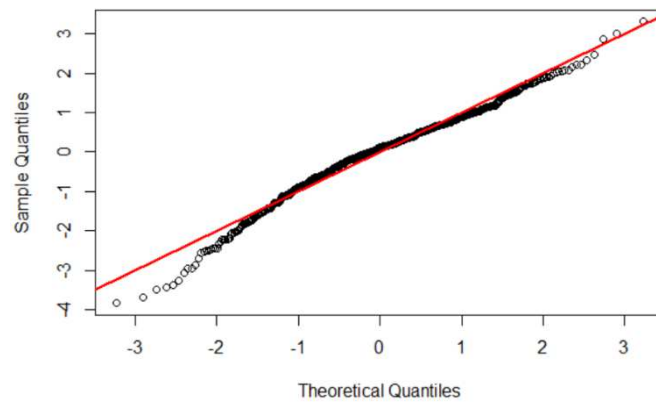
Figure 2 - Frequency histogram of the scores obtained by unhealthy companies



Source: our elaboration

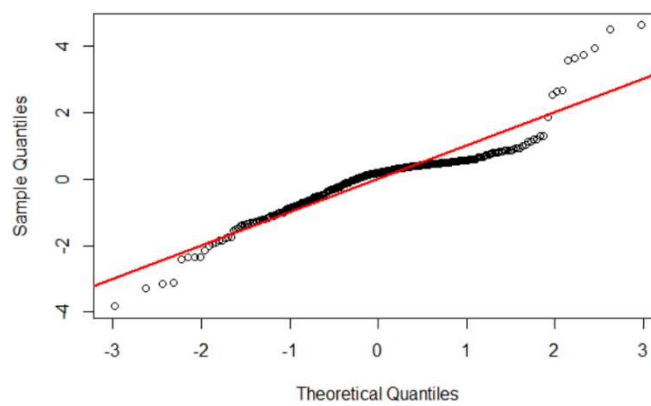
Normal Q-Q plots are depicted in Figure 3 and Figure 4. They are the most powerful tool to assess normality (Yap & Sim, 2011). In the specific case, the scores' distribution for the group G_1 is more skewed and more heavy-tailed than the one for the other cluster, which seems to be close to a Gaussian.

Figure 3 - Q-Q plot for good companies' scores



Source: our elaboration

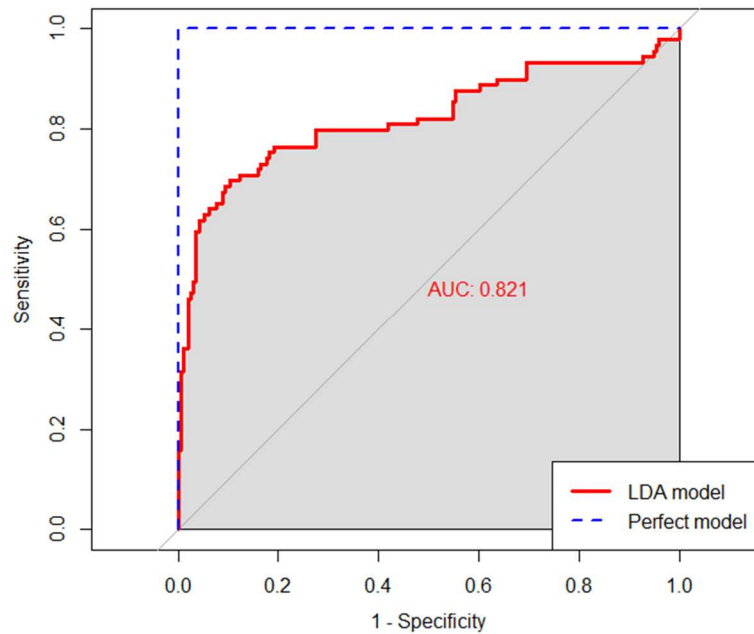
Figure 4 - Q-Q plot for bad companies' scores



Source: our elaboration

Using the test set, the next step consists of showing the ROC curve and computing its AUC value³. The latter, which can be interpreted as the likelihood of correct classification (Witzany, 2017), confirms that the model has a good discriminative power.

Figure 5 - ROC curve for fitted LDA model



Source: our elaboration

Once the scores have been computed, it is possible to calculate the PD, using the formula in Resti & Sironi (2008). The authors show also that, in order to assess if a company is defaulted (thus not reliable), a cut-off value must be calculated, and it is the simple average between the mean scores of each group (this will be called “solution 1”). Alternatively, it is possible to set a threshold for the PD, such that the bank should reject the loan request if the estimated PD is greater; with this knowledge, the cut-off value can be calculated considering this threshold (this will be called “solution 2”). Assuming the threshold to be equal to 50%, Table 3 shows the proportion of firms that should receive the money, according to each of the two solutions for the cut-off.

³ For details about ROC and AUC, see Witzany (2017) for example.

Table 3 - Proportion of borrowers for each solution

Cut-off	Proportion of borrowers (%)
Cut-off from solution 1	62.51
Cut-off from solution 2	79.23

Source: our elaboration

The latter two values assert that with the second cut-off about 17% more companies should receive the money from the bank, but this means that the credit institute should lend the money even to some firms that are currently already defaulted: in fact, the original proportion of non-defaulters is equal to 68%.

4 The logistic regression

4.1 Fitting the model

The sample dataset and the predictors are the same that has been utilised in the context of the discriminant analysis, since the purpose is to do a comparison between both models, and this would not make sense if datasets were different.

Table 4 illustrates the estimated coefficients obtained after having fitted the model.

Table 4 - Estimated coefficients from fitted logistic regression model (1)

Coefficient	Estimate	Standard error	p-value
Intercept	-1.08470	0.10367	< 2e-16
ROA	-1.51684	0.25120	1.56e-09
Quick ratio	-0.29234	0.08092	0.000303
Fixed assets coverage ratio	-0.39302	0.09072	1.48e-05
Indebtedness ratio	0.80359	0.63446	0.205309
Asset turnover ratio	-0.72629	0.21203	0.000614
Current asset turnover ratio	-0.89909	0.22490	6.39e-05
Bank debts on current assets	0.71038	0.25706	0.005719
ROD	-0.79658	0.72124	0.269396

Source: our elaboration

The coefficients for the two variables “Indebtedness Ratio” and “ROD” are not statistically significant with the confidence level of 5%, so they can be removed and then the analysis is repeated. At this point, all the regressors are statistically significant.

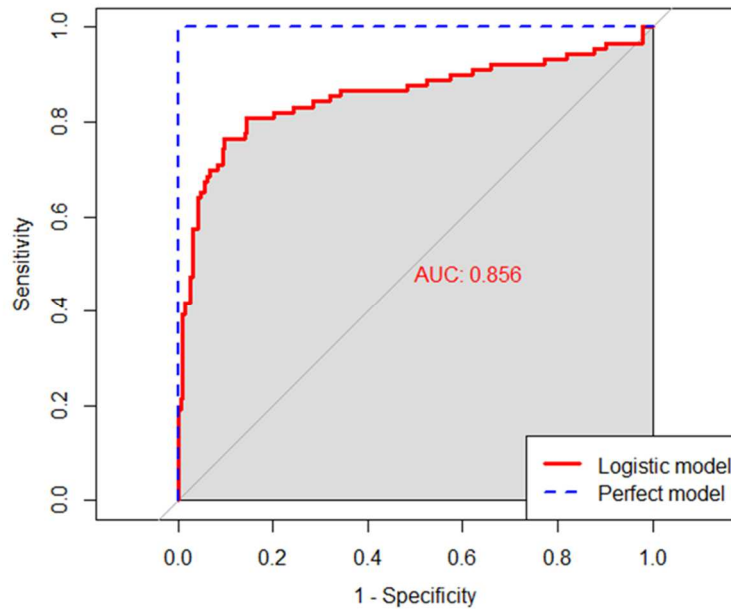
Table 5 - Estimated coefficients from the fitted logistic regression model (2)

Coefficient	Estimate	Standard error	p-value
Intercept	-1.11774	0.09265	< 2e-16
ROA	-1.37046	0.23239	3.69e-09
Quick ratio	-0.29543	0.08045	0.000241
Fixed assets coverage ratio	-0.37391	0.08746	1.91e-05
Asset turnover ratio	-0.68960	0.21201	0.001143
Current asset turnover ratio	-0.96671	0.22406	1.60e-05
Bank debts on current assets	0.67841	0.25856	0.008695

Source: our elaboration

These values are obtained once the two variables have been excluded from the analysis

The ROC curve related to the reduced model is showed in the next figure. The fitted logistic model is better than the previous one, since the AUC value is greater.

Figure 6 - ROC curve for fitted logistic model

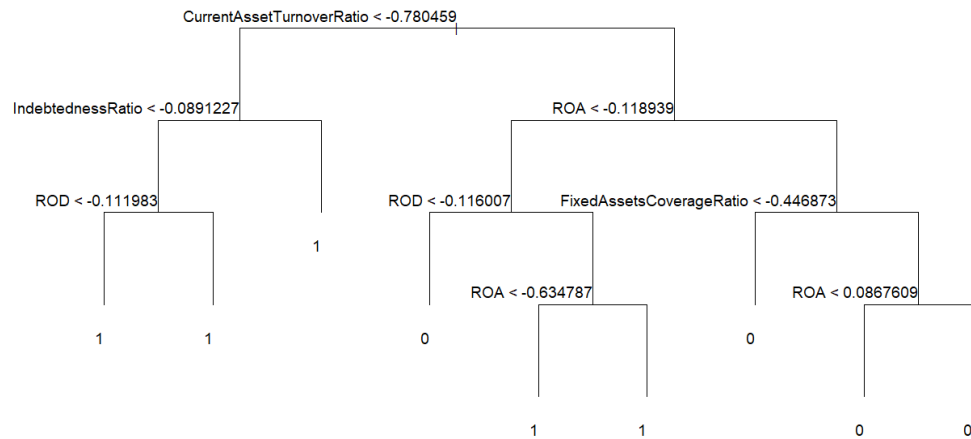
Source: our elaboration

5 The classification tree

5.1 Fitting the model

The data sample is the same of previous paragraphs and includes only those variables that have been selected to build the discriminant analysis model.

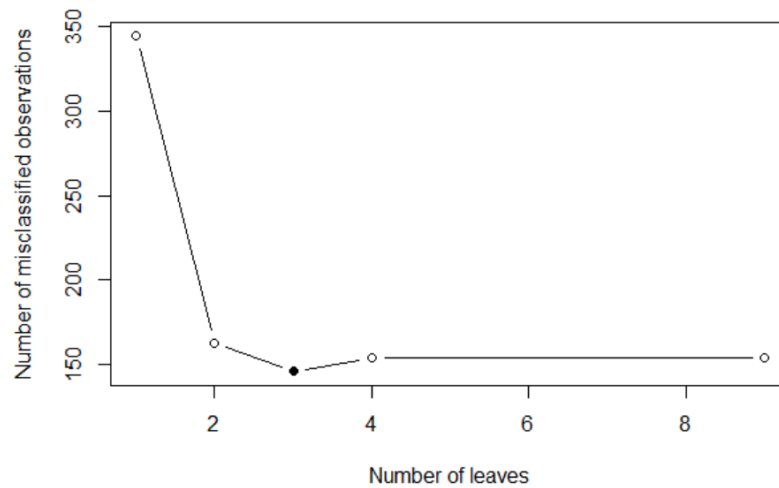
Figure 7 is interpreted as follows: each branch on the left of a node indicates that the splitting condition of the node is true, whilst the opposite branch indicates that the condition is false. This figure explains what the most important variables are: the one that best divides the observations is “Current assets turnover ratio”, so, if a firm has a very low value for that variable, the default will be almost certain.

Figure 7 - The fitted classification tree

Source: our elaboration

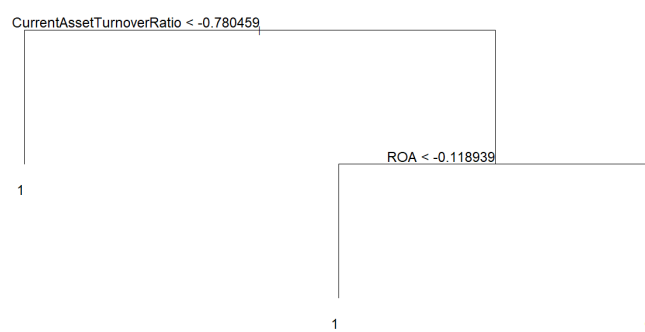
An important fact has to be noticed: several terminal nodes have the same predicted value, even though they have the same parent node. Thus, why the split is performed? It is performed because it increases the node purity: in fact, despite the same predicted value, the PDs change from node to node (James et al., 2013).

In order to prune the tree, it is necessary to make a plot of the number of leaves (that remains after the pruning) against the number of misclassified observations: this represents an impurity function (Giudici, 2005). As Figure 8 shows, the number of misclassified units is minimum when the tree has 3 leaves.

Figure 8 - The impurity function

Source: our elaboration

Figure 9 is very straightforward: a company is likely to default either if its value for the variable “Current assets turnover ratio” is smaller than -0.78 or if the “ROA” is smaller than -0.12 when the “Current assets turnover ratio” is greater than -0.78.

Figure 9 - The pruned tree

Source: our elaboration

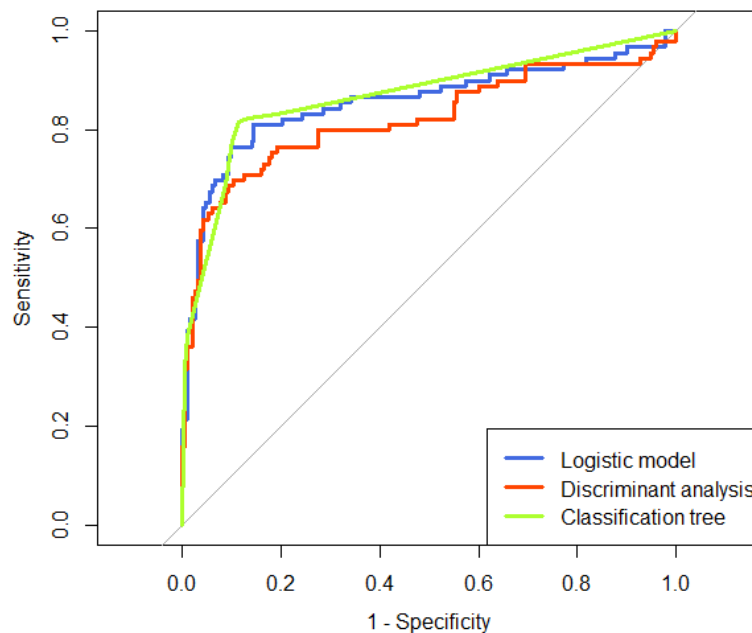
5.2 Comparing the Three Models

The comparison is possible thanks to Figure 10, where the ROC curve for the fitted LDA model is below the others, so this is the worst model in this case. On the other hand, the curve representing the fitted tree is the closest to the theoretical perfect model; overall, it is anywhere above the other two curves, except for a small range of values. Table 6 confirms that, in fact the AUC value related to the fitted tree is the highest.

It is important to notice that the ROC curves for the classification tree and the logistic regression (and their relative AUC) are really close to each other, so these two fitted models give basically the same performance. However, the classification tree approach is simpler, more intuitive and easier to communicate than logistic regression (Katsikopoulos et al., 2018; Aikman et al., 2014).

To sum up, going back to the initial question, it is possible to say that the fitted classification tree is the most reliable model, so bank's management should choose this one when evaluating a request from a company of the Italian food industry.

Figure 10 - Comparison of models



Source: our elaboration

Table 6 - AUC for each model

Model	AUC
Logistic regression	0.8556209
LDA	0.8208069
Classification Tree	0.8691761

Source: our elaboration

6 Conclusions

This article has shown that, although the differences in performance are little, the classification tree is the best model between the three analyzed, but this does not mean that classification trees are always better than the other two approaches: two reasons can be mentioned at least.

Firstly, the performance depends on the nature of the problem, specifically on the kind of relationship amongst the variables: if the variables have a linear relationship, a linear model (such as LDA or logistic regression) should be preferred; if the relationship is non-linear, building a decision tree could be the best option to correctly fit the data (James et al., 2013).

Secondly, there may be sample selection bias: this article has analyzed only the Italian food industry, thus credit institutes might not find this specific fitted model useful when assessing the reliability of companies belonging to other industries and/or other countries. However, Witzany (2017) mentions that Altman's Z-Score (Altman, 1968) has been criticized in many studies because of its potential sample selection bias, but it is still a useful tool in banking practice.

The approaches explained in this paper are the simplest. There are others more complex, but complexity comes with a cost: although model's performance may increase with complexity, more assumptions could be required, so the model could lose a part of its predictive capacity and could become less close to reality and less transparent. This trade-off should be considered by the bank's management, who should opt for simpler models (Aikman et al., 2014; Katsikopoulos et al., 2018). There is another concern to be considered, that is the bias-variance⁴ trade-off: if the complexity increases, bias will decrease but variance will increase (Katsikopoulos et al., 2018). However, variance seems to be more critical than bias (Brighton & Gigerenzer, 2015). In particular, Katsikopoulos et al. (2018) point out that simple models outperform more complex models if only poor quality information

⁴ For the definitions of "bias" and "variance", see Katsikopoulos et al. (2018) for example.

is available or it is not quantitatively sufficient: this problem may arise when the future is unpredictable. It seems that this may be the case: in fact, precise information about the default event was not available, and the current pandemic unleashed by SARS-CoV-2 has made the future more uncertain. Therefore, the management should choose simpler models, i.e. the classification tree in this case.

Finally, it is very important to remember that data were referred to a period before the current economic crisis, so analyzing the differences in the PDs before and after the current downturn may be the topic of a future article.

References

- Adnan Aziz, M., & Dar Humayon, A. (2006). Predicting corporate bankruptcy: Where we stand? *Corporate Governance: The International Journal of Business in Society*, 6(1), 18-33. <https://doi.org/10.1108/14720700610649436>
- Aikman, D., Galesic, M., Gigerenzer, G., Kapadia, S., Katsikopoulos, K., Kothiyal, A., Murphy, E., & Neumann, T. (2014). Taking Uncertainty Seriously: Simplicity versus Complexity in Financial Regulation. *Bank of England Financial Stability Paper*, 28. <https://doi.org/10.2139/ssrn.2432137>
- Alharan, A., Alsagheer, R., & Al-Haboobi, A. (2017). Popular Decision Tree Algorithms of Data Mining Techniques: A Review. *International Journal of Computer Science and Mobile Computing*, 6, 133-142.
- Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589-609.
- Badarau, C., & Lapteacru, I. (2020). Bank risk, competition and bank connectedness with firms: A literature review. *Research in International Business and Finance*, 51, 101017. <https://doi.org/10.1016/j.ribaf.2019.03.004>
- Biggeri, L., Bini, M., Coli, A., Grassini, L., & Maltagliati, M. (2012). *Statistica per le decisioni aziendali*. Pearson.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.
- Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, 68(8), 1772-1784.
- Cerved. (n.d.). *Note Metodologiche*. https://www.cerved.com/pub/bilanci/note_1_trend.pdf
- Coolen-Maturi, T., & Coolen, F. (2019). Non-parametric predictive inference for the validation of credit rating systems. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4), 1189-1204.
- Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-232.
- Regio decreto 16 marzo 1942, n. 267, "Disciplina del fallimento, del concordato preventivo, dell'amministrazione controllata e della liquidazione coatta amministrativa" (1942, 6 aprile) (Italia). *Gazzetta Ufficiale*, (81), 24-56. <https://www.gazzettaufficiale.it/eli/gu/1942/04/06/81/sg/pdf>

- Eisenbeis, R. A. (1977). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, 32(3), 875-900.
- European Commission. (n.d.). *What is an SME?* https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en
- European Communities. (2008). *NACE Rev. 2, Statistical Classification of Economic Activities in the European Community*. Office for Official Publications of the European Communities.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188.
- Giudici, P. (2005). *Data mining. Modelli informatici, statistici e applicazioni*. McGraw-Hill Companies.
- Giudici, P., & Figini, S. (2005). *Applied data mining: Statistical methods for business and industry*. John Wiley & Sons.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.
- Hunt, B. (1962). *Concept Learning: An Information Processing Problem*. John Wiley & Sons.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
- Jiao, S., Song, J., & Liu, B. (2020). A Review of Decision Tree Classification Algorithms for Continuous Variables. *Journal of Physics: Conference Series*, 1651(1). <https://iopscience.iop.org/article/10.1088/1742-6596/1651/1/012083/pdf>
- Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(2), 119-127.
- Katsikopoulos, K. V., Durbach, I. N., & Stewart, T. J. (2018). When should we use simple decision models? A synthesis of various research strands. *Omega*, 81, 17-25. <https://doi.org/10.1016/j.omega.2017.09.005>
- Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96(454), 589-604.
- Kim, K. S., Choi, H. H., Moon, C. S., & Mun, C. W. (2011). Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics*, 11(3), 740-745.
- Loh, W.-Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, 3, 1710-1737.
- Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, 82(3), 329-348.
- Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815-840.
- Loh, W.-Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, 83(403), 715-725.

- Messenger, R., & Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67(340), 768-772.
- Mihalovič, M. (2016). Performance Comparison of Multiple Discriminant Analysis and Logit Models in Bankruptcy Prediction. *Economics & Sociology*, 9, 101-118. <https://doi.org/10.14254/2071-789X.2016/9-4/6>
- Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302), 415-434.
- Naili, M., & Lahrichi, Y. (2020). The determinants of banks' credit risk: Review of the literature and future research agenda. *International Journal of Finance & Economics*, 1-27. <https://doi.org/10.1002/ijfe.2156>
- Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1), 109-131.
- Puri, M., Rocholl, J., & Steffen, S. (2011). On the Importance of Prior Relationships in Bank Loans to Retail Customers. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.1572673>
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- Resti, A., & Sironi, A. (2008). *Rischio e valore nelle banche. Risk management e capital allocation*. EGEA.
- Whitaker, T., Beranger, B., & Sisson, S. A. (2019). Logistic regression models for aggregated data. *arXiv: Methodology*.
- Witzany, J. (2017). *Credit Risk Management*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-49800-3>
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., & others. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), 1-37.
- Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12), 2141-2155.
- Zamore, S., Ohene Djan, K., Alon, I., & Hobdari, B. (2018). Credit Risk Research: Review and Agenda. *Emerging Markets Finance and Trade*, 54, 811-835. <https://doi.org/10.1080/1540496X.2018.1433658?needAccess=true>
- Zavgren, C. V. (1985). Assessing the vulnerability to failure of American industrial firms: A logistic analysis. *Journal of Business Finance & Accounting*, 12(1), 19-45.