SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Dipartimento di Ingegneria informatica, automatica e gestionale ”Antonio Ruberti”
PhD in Engineering in Computer Science

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Will it fail and why?

A large case study of company default prediction with highly interpretable machine learning models

Thesis Advisor
Prof. Aris Anagnostopoulos

Candidate
Stefano Piersanti
668719

Academic Year MMXXI-MMXXII (XXXIII cycle)

a Daniela e ai miei tre tesori

Abstract

Finding a model to predict the default of a firm is a well-known topic over the financial and data science community.

Default prediction problem has been studied for over fifty years, but remain a very hard task even today. Since it maintains a remarkable practical relevance, we try to put in practice our efforts in order to obtain the maximum prediction results, also in comparison with the reference literature.

In our work we use in combination three large and important datasets in order to investigate both bankruptcy and bank default: a state of difficulty for companies that often anticipates actual bankruptcy. We combine one dataset from the Italian Central Credit Register of the Bank of Italy, one from balance sheet information related to Italian firms, and information from AnaCredit dataset, a novel source of credit data by European Central Bank.

We try to go beyond the academic study and to show how our model, based on some promising machine learning algorithms, outperforms the current default predictions made by credit institutions. At the same time, we try to provide insights on the reasons that lead to a particular outcome. In fact, many modern approaches try to find well-performing models to forecast the default of a company; those models often act like a black-box and don't give to financial institutions the fundamental explanations they need for their choices. This project aims to find a robust predictive model using a tree-based machine learning algorithm which flanked by a game-theoretic approach can provide sound explanations of the output of the model.

Finally, we dedicated a special effort to the analysis of predictions in highly unbalanced contexts. Imbalanced classes are a common problem in machine learning classification that typically is addressed by removing the imbalance in the training set. We conjecture that it is not always the best choice and propose the use of a slightly unbalanced training set, showing that this approach contributes to maximize the performance.

Keywords: bankruptcy, default, financial stability, machine learning, binary prediction, explainability, imbalanced scenario, performance indicators.

Contents

Chapter 1

Introduction

In this project we faced the firms bankruptcy prediction problem. It represents a hard problem but also a intriguing challenge. Bankruptcy prediction of a company has been studied in the literature for over fifty years; the first relevant study at the beginning of the history of this field of research date back to Altman in the year 1968, when he proposed its well-known Z-score. But the scientific production on the subject is still intense nowadays and the results in default prediction are still not completely satisfactory today.

It is worth mentioning that even today, after more than 50 years, Altman's Z-score is still used by many banks to perform credit scoring activity, albeit in a slightly updated form compared to the original version.

In more recent years default prediction problem was also addressed using Machine Learning (ML) algorithms which have joined the more traditional statistical techniques.

We have worked hard on this problem, measuring its inherent difficulties. Our approach involved the latest ML techniques, also in combination with each other, performing an extensive data pre-processing work. We used an impressively large database, that contains three different data sources, which combine information relating to the balance sheet data of the largest Italian companies with data relating to the credit exposures of that companies. Finally, we have dedicated a particularly effort to the explanation of the results achieved: a very relevant issue in default prediction problem. We have focused on bankruptcy in the classical sense and on the bank default of firms, representing the latter a state of difficulty for companies that often anticipates actual bankruptcy. It is less studied in the literature, also due the lower availability of information.

According to our knowledge it is the first time that both situations of difficulty of companies are considered in the same study.

Nowadays, data science is becoming more and more important to support banks and businesses in the process of decision making. The main stakeholder that we consider are Central banks, Supervisory Authorities and private banks in their activity of credit risk management. Also general Governments may be interested in some cases in assessing the soundness of firms, for example when they provide public guarantees to companies in order to foster the recovery of economic activity.

The analysis of the problem is a double-edge sword, on one hand banks try to minimize their losses looking for the best model that will predict if a company will pay back its loan or not before allowing it, on the other hand many company nowadays may have not access to credit because of an inappropriate model or predictions that does not rely on machine learning approaches.

It is important to remark that our reference context is inherently strongly unbalanced (i.e. the companies that fail are in a very small percentage compared to the total); this implies additional difficulties in the predictions and in the measurement of the results. A relevant part of the work is dedicated to addressing this aspect, well known in the literature, in order to develop a solid model aimed at maximizing the performance of default prediction.

To conclude this brief introduction, in Table 1.1 we report a selection of techniques used over the time with reference to bankruptcy prediction. In the following we will provide a more detailed description about the problem.

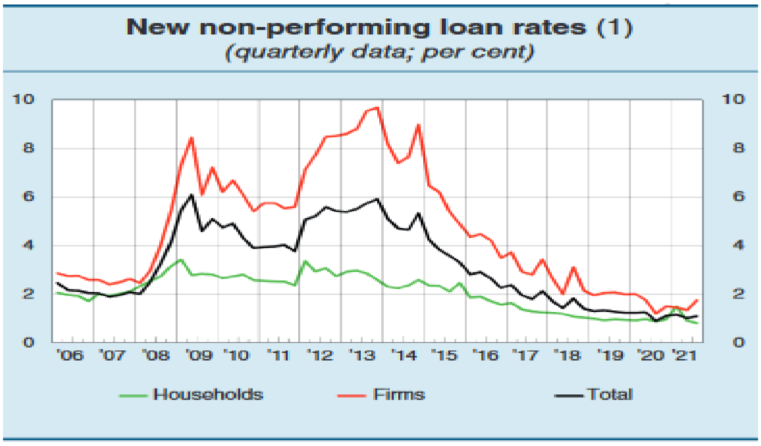
Bankruptcy decision techniques	
Statistical techniques	**Machine Learning (ML) techniques**
Linear Discriminant Analysis (LDA)	Artificial Neural Network (ANN)
Multi Discriminant Analysis (MDA)	Support Vector Machine (SVM)
Logistic Regression (LR)	Decision Tree (DT)
	Random Forest (RF)
	Other ML techniques

Table 1.1: The principal techniques used to address the bankruptcy prediction problem.

1.1 Description of bankruptcy prediction problem

Bankruptcy prediction of a company is, not surprisingly, a topic that has attracted a lot of research in the past decades by multiple disciplines [4, 6–8, 12–14, 18, 23, 27, 28, 37, 38, 47, 48]. The importance of such research stems from its financial applications in bank lending, investment, governmental support, and financial stability in general.

In particular, default prediction is one of the most challenging activities for managing credit risk. In fact, banks need to predict the possibility of default of a potential counterpart before they extend a loan. An effective predictive system can lead to a sounder and profitable lending decisions leading to significant savings for the banks and the companies and, most importantly, to a stable financial banking system. A stable and effective banking system is crucial for financial stability and economic recovery as well highlighted by the last global financial crisis started in 2008 and the next European debt crisis. Default prediction will also play a crucial role in the current economic situation characterized by the crisis generated by the Covid pandemic.



Source: Central Credit Register.
(1) Annualized quarterly flows of adjusted NPLs in relation to the stock of loans at the end of the previous quarter, net of adjusted NPLs. Data seasonally adjusted where necessary.

Figure 1.1: Dynamic of the new Non-Performing loans (NPLs) in Italy, over the years.

In Figure 1.1 we can observe the evolution of new Non-Performing loans (NPLs) in Italy over the years. NPLs represents a categories of loans that probably will determine a future loss in banks portfolio. The loan deterioration rate more than doubled after the first financial crisis in 2008, followed by the increasing in corporate bankruptcies.

More in general, the main reasons that motivate the relevance of bankruptcy prediction are excellently synthesized in [36] as follow:

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

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

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

- *Identification of sector wide problems*: Bankruptcy prediction models that flag firms belonging to a certain sector are likely to be a leading indicator of an upcoming downturn in a certain sector of an economy. With robust models, the business managers and government policy makers would become aware and take corrective action to limit the magnitude and intensity of the downturn in the specific sector. Industry groups in turn has been shown to significantly effect forecasting models (see Chava and Jarrow, 2004 [46]).

- *Signal to Investors*: Investors can make better and more informed decisions based on the prediction of bankruptcy models. This not only forces the management of firms to take corrective action but also helps to soften the overall economic fallout that results from the bankruptcies. Empirical studies have shown that investment opportunities are significantly related to likelihood of bankruptcy (see Lyandres and Zhdanov, 2007 [17]).
- *Relation to adjacent problems*: Bankruptcy prediction is often the first step used by ratings agencies to detect financial distress in firms. Based on the predictions of bankruptcy models, ratings agencies investigate and assess credit risk. Getting flagged by bankruptcy prediction models is often the first step that triggers the process of revising credit ratings. A literature survey covering 2000–2013 demonstrates the close relation between bankruptcy prediction and credit risk (see García et al., 2015 [21]).

The magnitude of bankruptcy costs is a critical issue in terms of capital structure theories. According to Fabio Panetta, former general director of the Bank of Italy, referring to Italian loans: *"The growth of the new deteriorated bank loans and the slowness of the judicial recovery procedures have determined a rapid increase in the stock of these assets, which in 2015 reached a peak of 200 billion, equal to 11 percent of total loans."* [1]

Several techniques for predicting bankruptcy have been developed over the years.

As shown in Table 1.1, statistical and machine Learning approaches are the two broad categories used to predict bankruptcy [6, 7, 23, 48].

Classical statistical techniques that have been employed include linear discriminant analysis (LDA), multi-discriminant analysis (MDA), and logistic regression (LOG), whereas Machine learning techniques (ML) include well-known algorithms such as Artificial neural networks (ANN), SVM, Decision trees (DT), Random Forest (RF) and Boosting techniques.

The main limitation of current research work can be attributed to the lack of data. The main source of data to predict whether a company will default or bankrupt is its financial statements reported in balance sheet. However such data are difficult for researchers to collect at a grand scale, as each company generally publishes its statements by its own means. Thus, past research has been based on small datasets of the order of a few tens of thousands companies.

Even worse, the financial statements do not capture the attitude of a firm towards its lenders. Past adversities of a firm to satisfy its debt requirements can be potentially very useful for predicting future difficulties. Of course knowledge of past behavior is even much harder to obtain than financial data.

[1] Fabio Panetta, Chamber of Deputies, Rome, May 10, 2018.

In this project we work on the problem of prediction of bankruptcy and of default of companies. In particular, we use historic data for predicting whether a company will enter in default.

We base our analysis on the use of three different datasets (see Chapter 2). First, we use historic information from all the loans obtained by almost all the companies based in Italy. This information includes information on the companies credit dynamics in the past years, as well as past information on relations with banks and values of protections associated with loans. The first dataset is based on information from Italian Central Credit Register which represents a historical Italian database containing information on the banking behaviour of companies and is widely used both by banks to assess the creditworthiness of counterparties and by banking supervisory authorities.

The second dataset that we use, also in combination with the first, contains a set of balance sheet indicator related to the Italian firms. Such data are typically used in the literature to predict corporate failure (see [7]) and can also significantly improve bank default predictions (see for example: Aliaj et al. [3]).

The third dataset contains information from a new credit data collection recently carried out by the European Central Bank (ECB), together with the National Central Banks (NCBs) of the euro area. This survey is called AnaCredit and is conducted in Italy by the Bank of Italy which collects important credit information on Italian companies from Italian banks, partly overlapping with those of the Central Credit Register and partly completely new.

In any case, the information collected in AnaCredit allows for an even greater degree of detail than those of the Central Credit Register.

We found that these new features are very important in default predictions. Due to the very recent period of use of the data, it is possible to combine credit information from AnaCredit dataset with the other source of information only for the more recent years.

Note that the dimensions and the information in our dataset are very significant also in comparison those of past work [6, 7], allowing to obtain a very accurate picture of the possibility to predict over various economic sectors.

In our knowledge, this surpasses by far any dataset that have been used for bankruptcy/default prediction, allowing to obtain a very accurate picture of the possibility to predict over various economic sectors.

In addition, we performed extensive data engineering and analysis on these data and we compare a variety of methods, achieving very high quality results.

However, for financial applications, classification accuracy does not suffice. The explainability of a model is fundamental when it is applied to financial problems because it is crucial that an interested stakeholder can comprehend the main drivers of a model-driven decision. Although not strictly connected with firms default, also the European GDPR EU (2016) regulation states that *"the existence of automated decision-making should carry meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."* In other words, under the GDPR regulation, the data subject is therefore, under certain circumstances, entitled to receive meaningful information about the logic of automated decision-making.

The focus of this project is on the empirical approach, especially the use of tree-based Machine Learning techniques. We will demonstrate that some of the more sophisticated boosting techniques achieve the best results when applied to this problem. Furthermore, we focus even more on what are the motivation that lead to a certain prediction. So, the accountability of the decision represents a crucial point in many scenarios. In fact, the explainability of a model is fundamental when it is applied to financial problems and for this reason, we used a modern approach called Shap that will show us the more relevant attributes that influence the outcome both for the default predictions regarding the entire dataset and even for each single prediction.

Of course, despite the plethora of studies over the past fifty years, predicting the failure of a company remain a very hard task.

Our conjecture is that part of the limitations in prediction is unavoidable and intrinsically connected with the data we use. In fact, the balance sheet data is a very accurate snapshot of a company's condition but is typically available with a long delay and therefore we always use data that is a little too old for predictions. At the same time, credit information can be available with lesser delay and therefore can be better used in a predictive key. But in some particular cases banks can continue to finance even companies in bad conditions (so-called *zombie lending phenomenon*) making credit information not useful.

Most related research has focused on bankruptcy prediction, which takes place when the company officially has the status of being unable to pay its debts (see Section 1.2). However, companies often signal much earlier their financial problems towards the banking system by going in default. Informally speaking, a company enters into a default state if it has failed to meet its requirement to repay its loans to the banks and it is very probable that it will not be able to meet his financial commitments in the future (again, see Section 1.2). Entering into a default state is a strong signal of a company's failure: typically banks do not finance a company into such a state and it is correlated with future bankruptcy.

We try to predict both bankruptcy and bank default using credit data in combination with balance sheet data. We will show that this combination of information will lead us to obtain very relevant results even in comparison with the best results in the literature. Also the use of particular feature selection techniques and the choice of some advanced boosting techniques play a crucial role in the final performance we obtained.

To summarize, in our work we face the problem of predicting corporate default, using a very large amount of data. Against this background:

- we propose to shift the focus on bank default which anticipates actual bankruptcy; this emphasize the usefulness of credit data in addition to balance-sheet data;

- we underline the need to accurately motivate and explain the predictions. In this regard the use of Shap is of considerable importance since it allows us to explain the main factors behind each single prediction. In fact, to explain a general model can be not enough and we show that every single prediction can be driven by different factors.

The need of prediction explainability is common to the main interested stakeholders which, in our opinion, fall into three main categories, namely

1. Private banks,

2. Supervisory Authorities,

3. Governments (when, for example, they provide public guarantees to companies).

The private banks do not have all the data that we used in the predictions, but the other two stakeholders we mentioned may have them.

However, first of all, the use of Shap for each single prediction can also be very useful also for private banks. In addition, we perform some experiments using only a small subset of our entire dataset (i.e. an amount of data that could be available by a medium private bank), showing that it is possible to obtain remarkable results in default prediction also in this case.

1.2 Definitions and problem statement

In our work we consider two main types of company failure. Most related research has focused on *bankruptcy* prediction, which takes place when the company officially has the status of being unable to carry out its business. It depends on the different jurisdictions, but it is generally the result of a judicial administrative procedure, which involves the decision of a court. However, companies often signal much earlier their financial problems towards the banking system by going in *banking default*.

Informally speaking, a company enters into a banking default state if it has failed to meet its requirement to repay its loans to the banks and, thus, it is very probable that it will not be able to meet his financial commitments in the future. Entering into a banking default state is a strong signal of a company's failure: typically banks do not finance a company into such a state and it is correlated with future bankruptcy. Firms bankruptcy prediction and more generally creditworthiness assessment of the companies can be very important also in *policy decisions*, such as for example the policies of assignment of public guarantee programs [5].

Next we provide more details about these two notions and the relationship between them. Even though the descriptions we provide here refer to the Italian (and to a large extent European) environment, the notions for other countries are similar.

1.2.1 Bankruptcy

The concept of *bankruptcy* of a business is the one most known to the public, and refers to the situation in which the firm ceases its business and it is unable to continue the production activity. It is in general in case the result of a legal finding. This information can be deduced from specific information sources that classify companies on an annual basis based on the condition recorded during each year. In particular we derive the information related to the activity state of each firms from our balance sheet dataset, that contains also information taken from the Italian business register (see Section 2.1). The main relevant situations are:

- *Active*: a firm that is registered as a regular operating company in the reference year.
- *Bankrupt (or failed)*: a firm registered as failed in the public register typically following a judicial sentence by a court that establishes the state of bankruptcy.
- *In insolvency*: the insolvency pertains to an objective situation of economic impotence, determined by the fact that the entrepreneur is unable to fulfill regularly, and with normal means, his obligations and the agreed deadlines.
- *In liquidation*: The termination of the company consists in the extinction of the productive combination following the definition of all the active and passive debt or credit relationships that belonged to the entity itself. Normally, the termination is preceded by a more or less long period which is identified in the liquidation phase.
- *Ceased or inactive*: It is no longer present in the register of companies (i.e. it appears in it as discontinued or inactive) as it has terminated its regular activity.

1.2.2 Adjusted default

The concept of banking default is often an early sign for a company's future failure. A firm is in *default* towards a bank, if it is unable to meet its legal obligations towards paying a loan. There are specific quantitative criteria that a bank may use to give a default status to a company.

The past financial crisis has led to a revision and harmonization at the international level of the concept of loan default and it has lead to the concept of adjusted default, which is the default concept that we consider in this paper.

The classification of *adjusted default status*, is a classification that the Italian National Central Bank (Bank of Italy) gives to a company that has a problematic debt situation towards the entire banking system. Naturally, also other National Central Banks have adopted a similar classification following EU harmonization. It represents a supervisory concept, whose aim is to extend the default credit status to all the loans of a borrower towards the entire financial system (banks, financial institutions, etc.). The term refers to the concept of the Basel II international accord of *default* of customers. According to this definition, a borrower is defined in default if its credit exposure has became significantly negative. To asses the status of adjusted default, the Italian National Central Bank considers three types of negative exposures, listed in increasing order of severity:

1. A loan is *past due* if it is not returned after a significant period past the deadline.
2. An *unlikely to pay* (UTP) loan is a loan for which the bank has high probability to loose money.
3. A *bad (performing) loan* is the most negative classification; *the likelihood of being paid is extremely low.*

Bank of Italy classifies a company in *adjusted default*, or *adjusted nonperforming loan* if it has a total amount of loans belonging to the aforementioned three categories exceeding certain pre-established proportionality thresholds [1]. depending to the severity of the impaired loan. Therefore, a firm's adjusted default classification derives from quantitative criteria and takes into account the company's debt exposure to the entire banking system. If a company enters into a status of adjusted default then it is typically unable to obtain new loans. Furthermore, such companies are multiple times more likely to bankrupt in the future.

The default status that we consider in this paper is the status of adjusted default. However, for brevity we may refer to is just as *default*.

1.2.3 Adjusted Default and Bankruptcy

To show the relationship between adjusted default and bankruptcy, we present in the following tables some statistics about the number of Italian companies that entered in adjusted default or bankrupted in recent years. Our starting point will be to consider the companies classified in adjusted default in 2015. In Table 1.2 we consider a broader concept that we call *No longer active firms* that include all the firms that for various reasons (see Section 1.2.1) are no more reported as *Active* in the public register from 2015 to 2019.

No longer active firms

	2015 (%)	2016 (%)	2017 (%)	2018 (%)	2019 (%)
adjusted default in 2015	10.4	19.4	24.9	27.5	31.3
no adjusted default in 2015	1.1	2.3	3.7	5.3	7.2

Table 1.2: Relationship between Adjusted default status for a company (in 2015) and a negative condition in the public register (from 2015 onwards). In particular we consider all the *No longer active firms* status, that include all the companies that are not reported as *Active* in the public register for the considered years. Among the Italian firms that entered in Adjusted default status in 2015, 10.4% were also *No longer active* in 2015, and 31.3% will be *No longer active* by 2019. On the other hand, only 1.1% of the firms not in Adjusted default in 2015 went *No longer active firms* in 2015, and only 7.2% by 2019.

Bankrupt firms

	2015 (%)	2016 (%)	2017 (%)	2018 (%)	2019 (%)
adjusted default 2015	3.0	4.9	8.1	10.3	11.7
no adjusted default 2015	0.1	0.2	0.6	1.0	1.6

Table 1.3: Relationship between Adjusted default status for a company (in 2015) and the *Bankrupt* condition in the public register (from 2015 onward). Among the Italian firms that entered in Adjusted default status in 2015, 3.0% were also Bankrupt in 2015, and 11.7% will be *Bankrupt* by 2019. On the other hand, 0.1% of the firms not in Adjusted default in 2015 went *Bankrupt* in 2015, and only 1.6% by 2019.

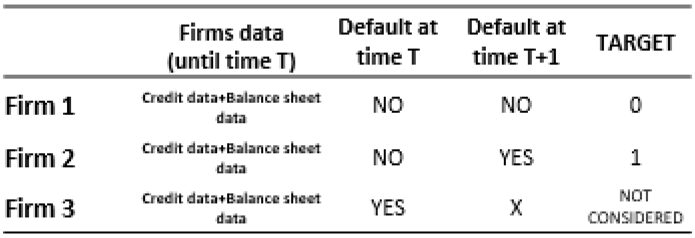
	Adjusted default firms				
	2015 (%)	2016 (%)	2017 (%)	2018 (%)	2019 (%)
% of total *No longer active* firms	34.9	32.8	29.9	27.6	21.2
% of total *Bankrupt* firms	60.2	62.1	56.6	53.4	51.0

Table 1.4: In the Table we analyze the link between the firms classified as Adjusted default in a particular year and the total number of firms in *Bankruptcy* (or *No longer active*) for the same year. For example we can observe that among the firms classified in Adjusted default in 2015 there are the 34.9% of *No longer active* firms and 60.2% of the *Bankrupt* firms. The percentage of *Bankrupt* firms identified as Adjusted default is higher than 50% for each considered year although slight decreasing over time.

1.2.4 Our target variables

Our goal in default prediction will be to classify whether a company will enter in adjusted default (and whether it will bankrupt) within the next year. According to our knowledge it is the first time that both situations of difficulty of companies are considered in the same study.

In the following Figure 1.2, we specify the conditions that identify our forecast target variables. In particular, we are interested in identifying companies that are not in adjusted default [bankruptcy] at time T and instead will be in adjusted default [bankruptcy] at time T+1 year. This will be our goal.



	Firms data (until time T)	Default at time T	Default at time T+1	TARGET
Firm 1	Credit data+Balance sheet data	NO	NO	0
Firm 2	Credit data+Balance sheet data	NO	YES	1
Firm 3	Credit data+Balance sheet data	YES	X	NOT CONSIDERED

Figure 1.2: In this table we specify our target variables. We use credit data and balance sheet data in order to predict the companies that are not in Adjusted default (or in Bankruptcy) at a particular time T and will be in Adjusted default (or in Bankruptcy) after one year.

1.3 A brief history

There has been an enormous amount of work on bankruptcy prediction. In order to give a flavor of how the literature that concerns bankruptcy prediction models has evolved, a brief review the most influential previous studies below is shown. We would like to provide a brief history about how it was dealt in the past and the most recent techniques.

1.3.1 Early Approaches: The Milestones

Initially, scholars focused on making a linear distinction among healthy companies and the ones that will eventually default. Among the most influencing pioneers in this field we can distinguish Altman [4] and Ohlson [38], both of whom made a traditional probabilistic econometric analysis. Altman, essentially defined a score, the Z discriminant score, which depends on several financial ratios (working capital/total assets, retained earnings/total assets, etc.) to asses the financial condition of a company. Ohlson on the other side, is using a linear regression logit model that estimates the probability of failure of a company and identifies four main factors that affect that probability: the company's size, its financial structure, its financial performance and its liquidity.

Some papers criticize these methods as unable to classify companies as viable or nonviable [8]. However, both approaches are used, in the majority of the literature, as a benchmark to evaluate more sophisticated methods. Those sophisticated methods are the machine learning techniques which are the focus of our project. Below, we provide a glimpse at the evolution of different techniques and at the comparison among them in the literature.

Since these early works there has been a large number of works based on machine-learning techniques [29, 35, 43]. The most successful have been based on decision trees [20, 28, 31, 49] and neural networks [6, 9, 18, 37, 48]. Typically, all these works use different datasets and different sets of features, depending on the dataset.

One of the first applications of the neural network methods for bankruptcy prediction, is that of Odom and Sharda [37]. They compared a three perceptron network against a method that was the "rule" until then: the multivariate discriminant analysis using the Altman ratios explained above, and it proved to be more robust in terms of accuracy. Boritz et al [9] uses two different techniques to train a neural network: back-propagation and Optimal Estimation Theory (OET), and compares them to more traditional methods such as discriminant analysis, probit and logit, as well as against benchmarks provided by directly applying the bankruptcy prediction models developed by Altman and Ohlson which were previously explained. They find no superiority for neural networks; instead the performance of each technique can be improved based on the relative size of the test and train set on the nature and number of imports etc. Nevertheless, it should be mentioned that they test only one architecture of a neural network. Atiya [6] suggests the inclusion of features extracted from equity markets because as he states "tend to be highly predictive, not only of the health of a firm, but also of the health of the economy, which in turn affects the creditworthiness of the firm", with a resulting improvement in accuracy of around four percentage points. What we can grasp from this retrospective look of previous experiments is that machine learning techniques probably will perform better in our bankruptcy prediction problem, but regarding the comparison among decision trees and NNs the feelings are mixed. We cannot, though, state with certainty ex-ante which of them will be superior.

There exists a wide range of classification methods included in the category of decision trees. Lee [27] by making a comparison of three of them using a dataset of Taiwan listed electronic companies concludes that the most efficient (in terms of accuracy) is the Generic Programming decision tree classifier which represents "a technique that applies the Darwinian theory of evolution to develop efficient computer programs". Zhou and Wang [49] on the other side, starting from the traditional random forest, propose the assignment of weights to each of the decision trees created, which are retrieved from each tree's past performance (out-of-bag errors in training method). They show that this modification improves the performance of the algorithm in terms of overall accuracy as well as the accuracy of their balanced dataset. Gepp and Kumar [20] turn their attention to another method, the Cox survival analysis, and compare it to the CART decision tree classifier and conclude that the former is the best one-year bankruptcy predictor, while the latter outperformed in the three-year prediction. Fernandez and Olmeda [18] compared NN with MDA, LR, MARS and C4.5 (two well-known methods that are based on the CART decision tree algorithm) on Spanish banks and showed that NN resulted in a higher accuracy. Martinelli et al. [31] by doing a very similar analysis on a database of Brazilian firms showed that the C4.5 algorithm is the one that outperforms the other methods.

1.3.2 More recent works: Does Machine Learning perform better?

Chakraborty and Joseph (2017) [11] train a set of models to predict distress in financial institutions based on balance sheet items, finding that ML approaches generally outperform statistical models based on logistic regression. Specifically, the RDF (Resource Description Framework) data model allows for a marked increase in discriminatory power of about 10 percentage points in terms of the Area under the Receiver Operating Characteristic (AuROC) compared with the logit model. Using data on US household default on mortgages, also Fuster et al. (2018) [19] find that the RDF model generates more accurate predictions than the logit model, although the improvement is minimal and accounts for about 1.2 percentage points of AuROC. The authors argue that most of these gains result from the sophisticated functional form of the RDF model, which captures the complex relationships connecting different variables to default outcomes with greater discriminatory power. Albanesi and Vamossy (2019) [2] develop a model to predict consumer default based on deep learning (i.e. a combination of forecasts from deep neural network and gradient boosting) in environments with high-dimensional data (over 200 variables).

In a recent very important study, Barboza et al. [7] compare such techniques with support vector machines and ensemble methods showing that ensemble methods and random forests perform the best.

Recently, Andini et al. [5] have used data from Italian Central Credit Register to assess the creditworthiness of companies in order to propose an improvement in the effectiveness of the assignment policies of the public guarantee programs. In 2020, also Moscatelli and al. [34] use credit data from Italian Central Credit register in order to predict companies default reaching interesting results in performance prediction. The majority of the cited works typically try to predict bankruptcy of a company. Our goal is to predict both bankruptcy and bank default, after explored the related connections between the two critical situations. Furthermore, most of these papers use balance sheet data (which are public). Our dataset contains a very granular information of a very large set of companies on the past behavior of loan repayment. In particular, we use two different important sources of credit information: the Italian Central Credit Register and, for the fist time given the novelty, the Italian component of AnaCredit, that is a very recent European credit dataset. We combine credit data with balance sheet data in order to show that the combination of the two source of information improve the prediction performances. To our knowledge, our dataset is one of the most extensive dataset used in the literature. But the crucial point we wanted to address is trying to explain the predictions, investigating through the use of Shap, the crucial factors that could explain the failures of companies.

1.3.3 Works on explainability of credit prediction

In the most latest years the focus of scholars started to migrate form the resarch of a well performing model to one with a meaningful explainability. Paolo Giudici et al. [39] tried to augment traditional credit scoring methods of peer to peer lending agents with "alternative data" that consist of centrality measures derived from similarity networks among borrowers in order to improve predictive accuracy as well as model explainability. Rudrani Sharma, Christoph Schommer, Nicolas Vivarelli [44] instead tried to build up explainability of credit risk models to some degree in trained MLPs through sensitivity analysis. Branka Hadji Misheva et al. [10] applied different explainability methods such as LIME, Shap to machine learning (ML)-based credit scoring models applied to the open-access data set offered by the US-based P2P Lending Platform, Lending Club. All this researches show up the growing interest in EXplainable Artificial Intelligence (XAI) among the scholar community that is building up strong fundamentals for the financial system of the coming years.

1.4 Contributions

To summarize the contributions of our work are:

1. We analyze three very large datasets with highly granular data on the performance of each company in the past. In particular we use two credit datasets from the Italian Central Credit Register and the AnaCredit survey, also in combination (over $570K$ companies) and a balance-sheet dataset (over $700K$ companies).

2. We use these data to predict whether a company will default within a year, considering both bankruptcy and bank default prediction. We try to extract the most of the information using an accurate procedure of feature selection, in combination with some promising recent boosting techniques. We show that balance sheet data are very useful in order to predict bankruptcy while credit data perform better when we try to predict bank default. We prove that in both cases that combination of the two different source of data (credit data plus balance sheet data) improve the prediction performances.

3. We emphasize the performance of our approach and the possible economic gains using a comparison with some techniques actually used by banks in default prediction and by applying our method to some important practical applications.

4. Moreover, we give a particular relevance to the ability to explain the prediction using a very promising tool, Shap, in order to provide a method to give a sound explanation of the general model. But, we underline also the importance of an ad-hoc explanation for each single prediction.

5. Finally, we dedicated an additional effort in analyzing the specificity of the prediction in a highly unbalanced context. In particular we show that the use of a fully balanced training set does not represent in general the best choice and we propose the use of a slightly unbalanced training set in order to maximize the performance. Moreover, we show that the evaluation of the prediction performances is highly dependent from the performance indicators we use. We confirm, in line with a recent literature (see [42]), that some metrics like F1-score and MCC perform best in highly unbalanced contexts. And conjecture that a customized cost function may, in particular cases, be more suitable than the most common metrics used in the literature.

Roadmap. In Chapter 2 we describe the datasets we used and in Chapter 3 all the techniques and algorithms. We present in Chapter 4 our experimental results related to default prediction, followed by a in-depth analysis relating to the explainability (Chapter 5).

In Chapter 6 we provide our analysis related to the prediction in an imbalance scenario and in the following Chapter 7 some discussion and some practical application.

We draw our conclusion in Chapter 8.

Chapter 2

Datasets

2.1 Datasets definition

Our analysis is based on three datasets, which we now describe.

The first dataset is composed of credit information related to a large sample of Italian companies collected by the Italian Central Credit Register. The second dataset also concerns information on bank lending that comes from a new European Central Bank (ECB) survey called AnaCredit (stands for Analitical Credit dataset). The third dataset, instead, reports balance sheet data of a large sample of Italian companies.

In the following we enter more in the details.

2.1.1 Central Credit Register dataset

The first dataset is a very large and highly granular dataset of credit information about almost all the Italian companies belonging to the Italian Central Credit Register (CCR dataset). The Italian Central Credit Register is an information system on the debt of the customers of the banks and financial companies supervised by the Bank of Italy. It collects information on customers' borrowings from the financial intermediaries (banks and other financial corporations) and notifies to the banks of the risk position of each customer towards the entire banking system.

The Italian banks report on a monthly basis to the Bank of Italy the total amount of credit due from their customers: data information about loans of at least $30,000$ euros and non-performing loans (NPLs) of any amount. The Italian Central Credit Register has three main goals:

- to improve the process of assessing customer creditworthiness,
- to raise the quality of credit granted by the banks, and
- to strengthen the financial stability of the credit system.

In this project we use a large dataset obtained from CCR data. It contains almost 800K firms for each quarter from 2014 to 2020. For each company and each quarter in this period, the dataset contains 20 different attributes related to credit; the most important are shown in the left part of following Table 2.1. Some more dataset information are reported in Table 2.2.

ID	(1) CCR	ID	(2) AnaCredit
C1	Granted amount of loans	A1	Firms default status assessed by banks
C2	Used amount of loans	A2	Loan exposure
C3	Banks classification of firms	A3	Arrears
C4	Average amount of used loans	A4	Initial loan committment
C5	Overdraft	A5	Outstanding nominal amount
C6	Margins	A6	Off balance-sheet amount
C7	Past due (loans not returned)	A7	Amount of protection
C8	Amount of problematic loans	A8	Probability of default (PD)
C9	Amount of non-performing loans		
C10	Amount of loans protected by a collateral		
C11	Value of protection		
C12	Previous Adjusted default classification		

Table 2.1: Main attributes for CCR dataset (on the left) and AnaCredit dataset (on the right). CCR and AnaCredit dataset both have quarterly frequency. The CCR dataset contains 20 features while the AnaCredit dataset contains 14 features. In our experiments (Chapter 4) we observe the relevance in prediction of some of these attributes. In particular: $Margins$ (i.e Granted amount minus used amount) that can show signs of difficulty for the firm, $Overdraft$ and $Arrears$ that indicate high difficulty in debt repayments. Another relevant attribute is, not surprisingly, the probability of default assessed by banks.

2.1.2 The AnaCredit survey

The European AnaCredit database is a source of detailed information on individual bank loans in the Euro area. AnaCredit, which stands for Analytical Credit datasets, is a shared multipurpose database containing loan-by-loan information on credit extended by credit institutions to companies and other legal entities. On May 18, 2016 the governing council of the European central bank (ECB) adopted Regulation ECB/2016/13 on the collection of granular credit and credit risk data (AnaCredit) establishing Stage 1 of a shared database for the European System of Central Banks (ESCB) as of September 2018. The database contains a large amount of credit information, updated mostly on a monthly basis, based on harmonized concepts and definitions common to all participating countries. The AnaCredit survey has been designed with the goal of obtaining a complete picture of:

- the total credit exposure of the European banks;
- the total indebtedness of borrowers across all lenders.

The information collected consists of about 90 different attributes based on harmonized concepts and definitions and covers various aspects of the credit exposure.

Our AnaCredit dataset

In the right part of Table 2.1 we can see the main features that we use in our project to try to predict the default of companies. This information also includes some important vulnerability indicators assessed by banks, such as default status and probability of default of companies that have relationships with banks. Our AnaCredit dataset represents a selected subset of the Italian components of the European AnaCredit database; it contains quarterly information for over 900K Italian companies for the years 2018 − 2020 (2018 is the starting point for the collection of AnaCredit data). The detailed information about the dimensions of the dataset are reported in Table 2.2.

Credit datasets

	CCR						
	2014	2015	2016	2017	2018	2019	2020
Firms (Total)	791,603	764,607	740,106	723,196	709,937	696,977	746,764
Firms (Defaulted)	30,229	22,501	16,457	14,368	12,740	12,422	

	AnaCredit		
	2018	2019	2020
Firms (Total)	866,719	841,546	906,614

Table 2.2: Numbers of firms included in the CCR dataset (on the left) and AnaCredit dataset (on the right). We can observe that the rate of firms that have defaulted has practically been reduced by half from 2014 to 2019. The concept of bank default (adjusted default) that we use in the paper is derived from the CCR data, therefore we have reported this information only for this dataset.

2.1.3 Balance sheet dataset

Our third dataset (Balance) consists of the balance sheet data for more than 600K Italian firms. They are generally medium and large companies. In this paper we use balance sheet information for the years from 2014 to 2020, also in combination with the other two credit dataset. In particular we use Balance dataset in Section 4.2.1 and 4.2.2. The main features include those that regard the profitability of a company, such as return of equity (ROE) and return of assets (ROA); in addition, a very important balance sheet indicator is the firm's rating based on other balance sheet information, calculated by Cerved Group, an Italian company specialized on the manipulation of financial data from companies. We can refer to Table 2.3 for an overview of the main dataset attributes and Table 2.4 for an overview about the dataset dimensions. In our experiments we use a balance sheet dataset derived from a Cerved archive in use at the Bank of Italy for economic analyzes, but typically balance sheet data are public data and have been used extensively for bankruptcy prediction in the most

BALANCE SHEET DATASET

(3) BALANCE

ID	Description	ID	Description
X0	Rating	X11	Margin on revenues
X1	Revenues	X12	Operating costs/production value
X2	Employees	X13	MOL/production value
X3	Total added value	X14	MOL/operational added value
X4	MOL	X16	Liquidity
X5	Total asset	X17	Total equity/financial debt
X6	Working Capital	X18	Bank financial debt+ICS/financial debt
X7	Net Equity	X19	Financial expenses/MOL
X8	Return on equity (ROE)	X20	EBITDA/Net financial charges
X9	Return on investment (ROI)	X24	Current profit
X10	Return on asset (ROA)	X34	Net revenues/operating assets

Table 2.3: Main attributes for the balance sheet dataset (Balance). Our balance sheet dataset contains 35 different features, including a large number of well-known balance sheet indicators like, for example, ROE and ROA. Moreover, in the Balance dataset there is also a rating indicator based of balance sheet variables performed by an Italian company specialized in the sector (Cerved). Balance dataset contains data for over 600,000 Italian firms with annual frequency for years 2014 − 2020.

important reference literature (see for example [7] and references therein).

BALANCE SHEET DATASET

(3) BALANCE

	2014	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*
Firms (Total)	554743	568385	577840	589043	605504	609525	472280
Firms (Bankrupt)		*8933*	*4149*	*4731*	*3465*	*3855*	*2948*

Table 2.4: Numbers of firms included in Balance dataset. We can observe also in this case the sharp reduction regarding the failed firms from 2014 to 2019. In this case we refer to the concept of bankruptcy that we describe in Section 1.2.

2.1.4 Combined credit datasets

In our experiments, in order to predict firms failures (both bankruptcy and bank default status) we use also two combined datasets created using data from the three elementary datasets we have considered so far: CCR, AnaCredit and Balance datasets. More in details, we use:

- a combination between CCR dataset and AnaCredit dataset for the years 2018 − 2020 (AnaCredit data are not available before) that we called Mixed dataset; this dataset contains almost $600,000$ firms and has the characteristics that we can observe in Table 2.5;
- a combination between all the three source of information: CCR, AnaCredit and

BALANCE dataset for the years $2018-2020$. In this case, the combined dataset, that we called MERGED dataset, contains information for about 380,000 firms (see again Table 2.5).

We use the two combined datasets in order to predict both firms bankruptcy and firms adjusted default.

COMBINED DATASETS

	BOTH **(1) + (2)**	**MERGED** **(1) + (2) + (3)**
Firms (Total)	578878	379001
Firms (Defaulted)	*11208*	*8017*
Firms (Bankrupt)	–	*751*

Table 2.5: Numbers of firms included in the two combined datasets. MIXED dataset is composed by CCR dataset and ANACREDIT dataset, while MERGED dataset combines all the three dataset we consider in this paper: CCR dataset, ANACREDIT dataset and BALANCE dataset. The two combined datasets (MIXED and MERGED) include information relating to the years 2018 to 2020 due to the limited historical depth of ANACREDIT. The data on failed firms was not considered for the MIXED dataset as this information is present only for the companies that appear in the balance sheet dataset.

Other (categorical) features We add to our widest combined dataset, MERGED (that represents the most extensive dataset we considered in terms of number of features) also some further features related to each firm. These other features cannot be classified either among credit or balance sheet variables. In fact, these are features that represent some characteristics of companies such as geographic location or sector of economic activity. These characteristics are different from the others we use as they are intrinsic to each company, do not change over time and do not depend on the firms activity or the economic cycle. Therefore, we conceptually consider them separately from the two main sources of information we are using. we call these additional attributes "categorical" features and we add them to the two datasets that combine both credit and balance sheet information in order to try to improve the default predictions even further. In fact, another relevant point is that they are in all cases categorical variables.

A list of the categorical variables we used is present in the Table 2.6.

CATEGORICAL FEATURES	
ID	**Description**
C1	CAP (detailed geographical localization)
C2	PVAFF (geographical localization)
C3	SETCON (Economic activity sector)
C4	ATECO_CTP (detailed economic sectorization)
C5	RAMO (Further economic classification)

Table 2.6: List of some other features used in our prediction experiments. These represent some intrinsic firms characteristic that do not change over the time and do not fall either in the credit data or in the balance sheet data. We call these attributes "categorical" features.

2.2 Exploratory data analysis

To provide a clear and precise overview of the whole merged dataset used, it is a good practice analyze our dataset getting from them some statistics. This process, with the use of some visualization tools, is fundamental to understand properly some characteristics of the huge amount of companies that are considered. This early study may helps understand better the type of problem this work is currently facing and may provide some hints on its resolutions. This is even more important due to the fact that this work is based on three different elementary datasets very heterogeneous. Two of them analyze the problem by the same credit point of view, but with different variables, while last one with balance data is different story.

2.2.1 An exploration of the elementary datasets

In the following we give some details about the three elementary datasets and the final combined dataset. We will explore some relevant characteristic of the data that we use in the following in order to perform the predictions and the further analysis.

With reference to the three elementary datasets we explore the feature relevance using a tree based classifier (Random Forest). We measure the importance of the features in each dataset by the so called Gini Importance or Mean Decrease in Impurity (MDI), that calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportionally to the number of samples it splits. We consider the relevance respect to the Adjusted default (as target variable) for CCR and AnaCredit dataset and respect to the Bankruptcy target variable for the Balance dataset.

The results are shown below. In Table 2.1 we can observe a significant relevance for the feature C5 (overdraft) in the CCR dataset; features A3 and A8 (arrears and probability of default) dominate in the context of the AnaCredit dataset (Fig. 2.2), while there is a significant relevance for X0 (rating) when we consider the Balance dataset (Fig. 2.3).

These evidences are a first interesting learning with the most important variables that we will use in the predictions of the default and we recall these finding and we will deepen them in the rest of the work.

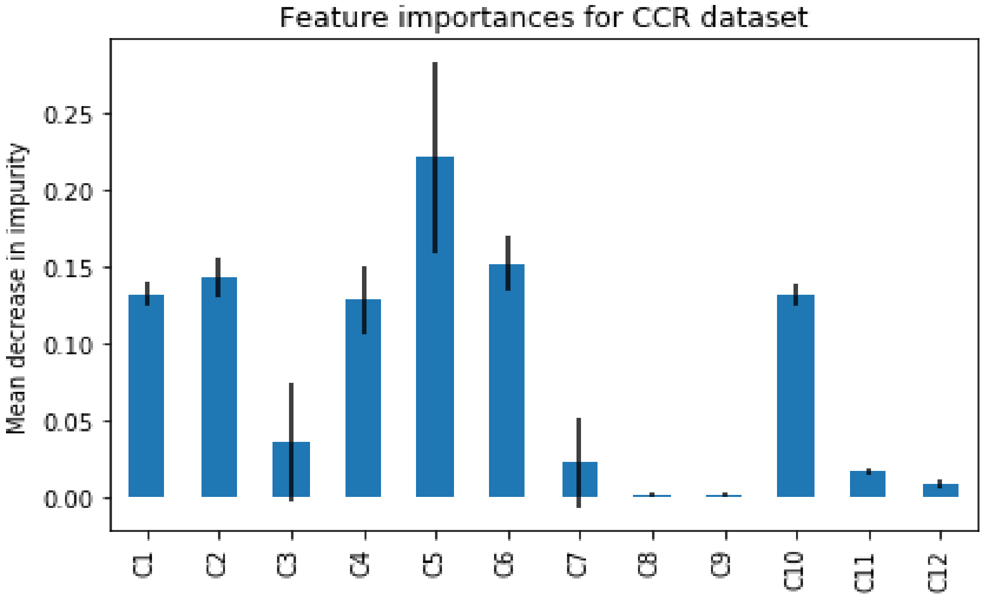


Figure 2.1: The Figure show the feature importances (in MDI) for the CCR dataset.

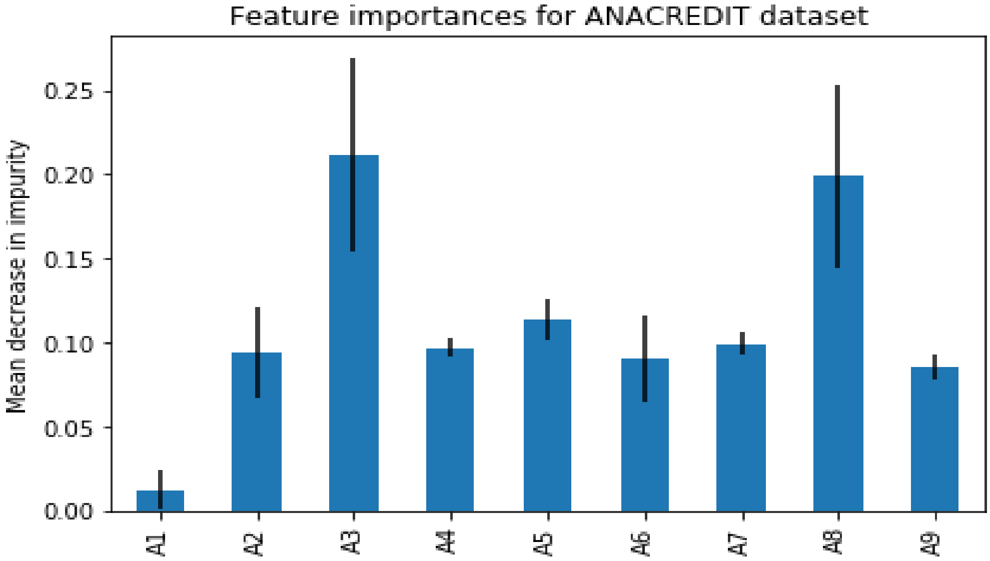


Figure 2.2: The Figure show the feature importances (in MDI) for the AnaCredit dataset.

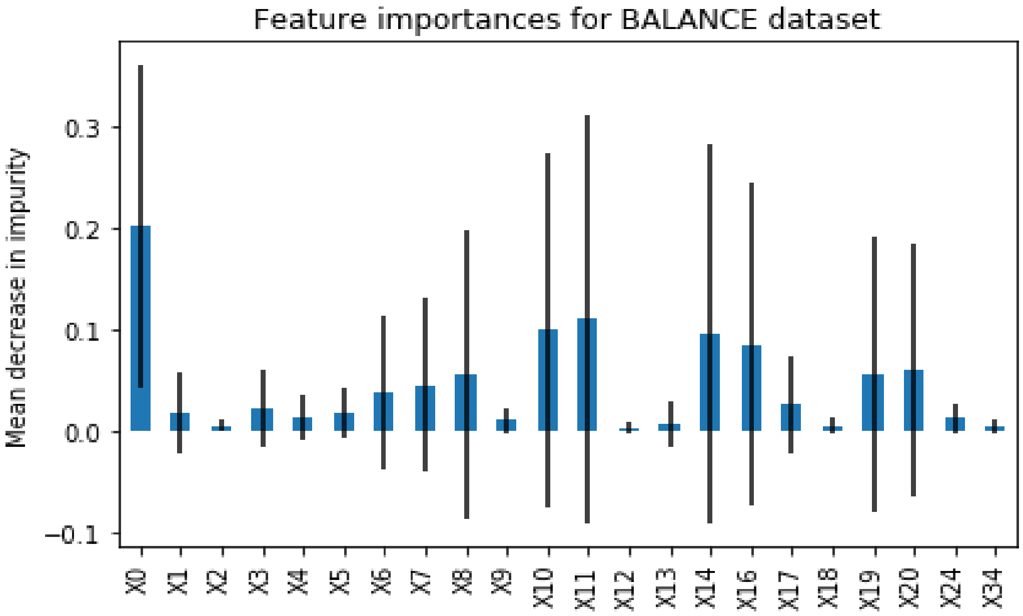


Figure 2.3: The Figure show the feature importances (in MDI) for the Balance dataset.

2.2.2 The combined dataset: a general overview

The three different datasets share the same companies in order to be merged in one big final dataset. As we mention before, we call this dataset as Merged dataset. The first really important thing to outline is the huge number of companies is: **three hundred seventy-nine thousand** different companies. At the moment it seems to be one of the biggest dataset in terms of numbers of companies observed, for a similar study. About it is important to observe which is the amount of them that after one year went into the state of adjusted default and how many of them went bankrupt. In Table 2.5 we show clearly that the problem is highly unbalanced, only the 2.1% of the companies in those datasets went on a state of adjusted default and less than 0.5% of them went bankrupt within a year.

2.2.3 A highly unbalanced dataset

From the definition of the problem was possible to expect this imbalance between the two classes but from a statistical point of view, it is a limitation. Unbalanced classifications is a real challenge for ML models since most of them when used for classification are designed around the assumption of an equal number of examples for each class. This may results in models that have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important, as it can clearly seen from our problem, and therefore there is the need to be more sensitive to classification errors for the minority class than the majority class. It is possible to overcome these types of problems by choosing carefully the evaluation criteria of the model and at the same time tweak the dataset preparation to minimize this problem as much as possible. There exist cases anyway, such the one in this work, in which the imbalance is a peculiarity of the problem itself and it is not cause by missing data from one class due to bad sampling, so one possible strategy base on a proper evaluation criteria may allow the model to be trained on an unbalanced dataset.

We will deepen this important issue in a following dedicated Chapter (Chap. 6).

2.2.4 Firms size dimension

Given this the next step that, in our opinion, deserved attention was to find a way to look for some relationship between the size of the companies and the outcome in terms of adjusted default. Unfortunately the information relating to the size of the company is not available with good quality for the whole sample of companies that we consider. We decide to use the average amount of loan granted to each firm in order to estimate the dimension of the companies. For this purpose it has been decided to divide the companies into three ranges that try to represent small, medium and large companies based on the total loans of the individual firms. The ranges does not belong to some specific definition but are the result of some experiments and in the end has been decided to set them as follows:

- Small: from 0 to 250,000;
- Medium: from 250,001 to 5,000,000;
- Large: above 5,000,000.

Our choice appear quite reasonable since we obtain an adequate breakdown of the dataset with a very similar number of medium-sized and small firms (about 180,000) while large firms are just over 22,000 (Fig. 2.4).

What can be observed is that there is some relevant changes in the ratio of failed companies between these three categories. In Figure 2.4 it is possible to see that there is some link between the amount of the loan, that is used to determine the dimension and the ratio of companies that undergoes into a state of adjusted default. Specifically large companies seems to be more able to repay the debt and it may be a sound result that may impact the future model. For the largest companies we can observe a 1.6% ratio for failed firms. Instead, between medium and small companies the difference becomes not relevant and we can be found a very similar ratio for the Adjusted default firms in these two categories (about 2.2%).

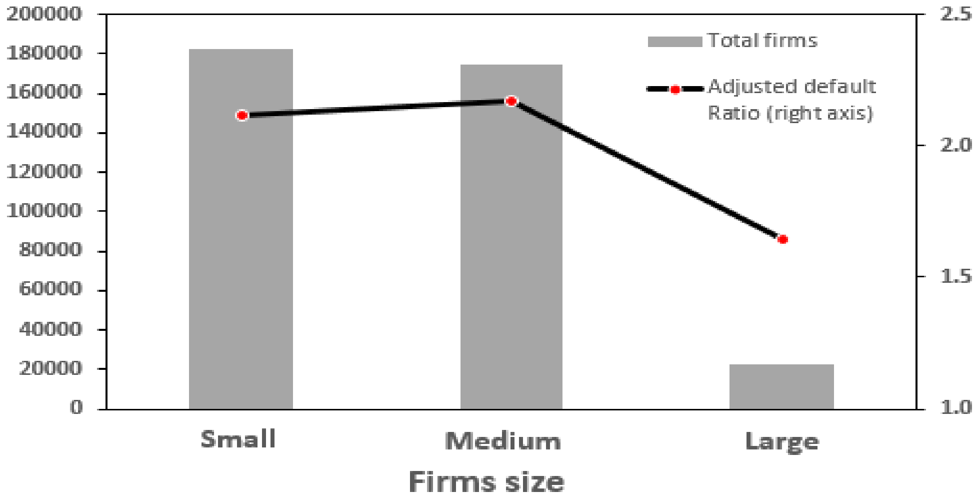


Figure 2.4: Adjusted default disentangle based on company dimension.

2.2.5 Categorical features

Inside this dataset there are interesting insights too, especially from high cardinality categorical variables. This time the selected variables relates more to the geographical position as well as the type of business the company runs. Those features are:

- ATECO: Code used to classify companies by the ISTAT based on the typology of business;
- PVAFF: affiliated province;
- SETCON: economic sector;
- CAP: the postal code;
- RAMO: business branch.

Each of this variable is a categorical data with high cardinality. In order to extract relevant information is important to understand if there are some categories for this variables in which the amount of failed company is higher or if they are evenly distributed. For this reason it is computed the ratio of companies in a status of adjusted default respect to the total in the same category. As it clearly possible to see in the following charts there is not a uniform distribution and this may be a useful information in the future while pre-processing those categories.

Geographical data Analyzing the ratio of companies in adjusted default based on the district and the postal code it is possible to notice the there are some area in which this effect has an higher magnitude. Analyzing by district in Fig. 2.5 there is a clear difference between zone. The first bar much different in high from the others represent all the companies that are base outside the national territory but are still considered inside the national economy.

Anyway this is a useful information that it will be considered in further processing. For what concerns instead the postal code, the analysis is much more granular due to the enormous amount of postal codes even in the same area (in Fig. 2.5 are showed just a small amount of them, the most important in terms of ratio of failed companies). Anyway is clear that some CAPs have an important ratio that cannot be ignored when dealing with our problem. It is important to say that going on with the list of caps ordered by ratio the effect becomes more and more similar.

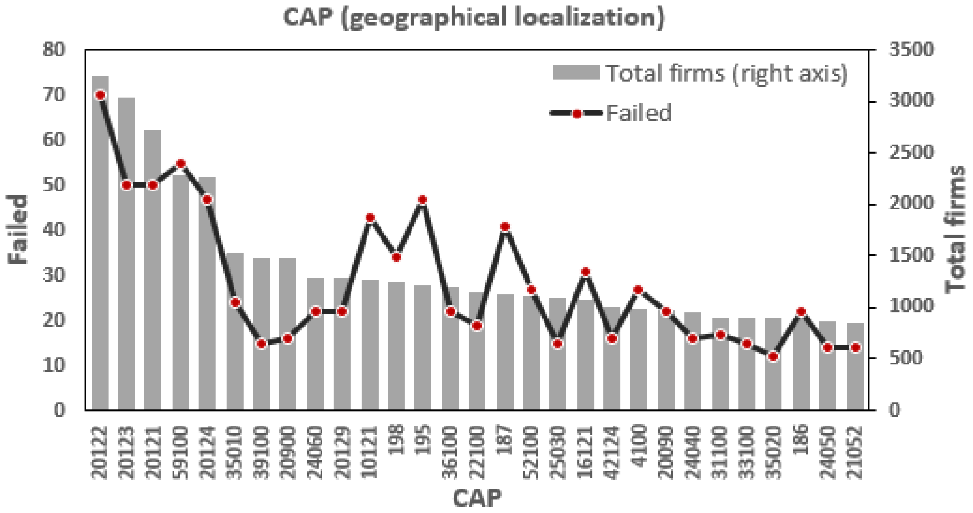


Figure 2.5: Ratio Adjusted default firms per CAP codes.

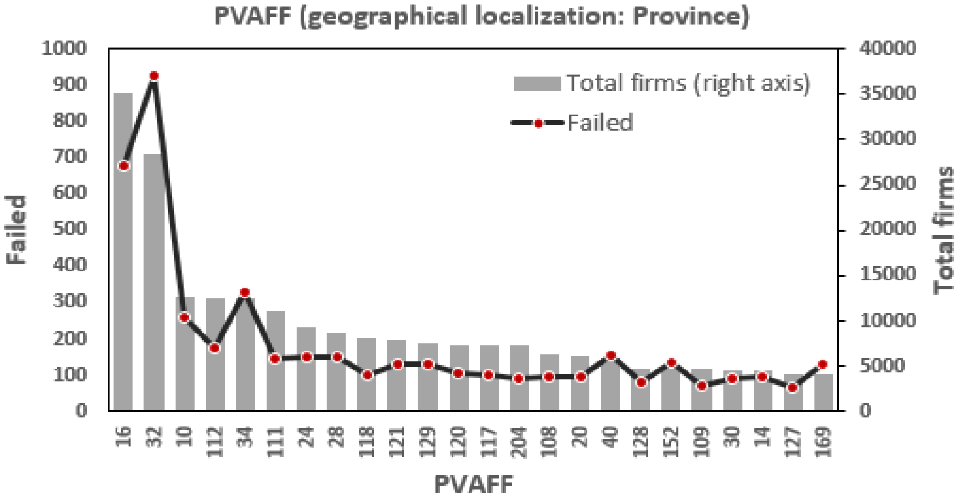


Figure 2.6: Ratio Adjusted default firms per Province codes.

Business data Dealing instead with data about the type of business what can be seen is that in terms of SETCON (Economic sector, Fig. 2.7) there are much less categories and there is one that is prevailing the others. That class is by definition the class of others activity and is interesting how the ratio is high in such class. When considering the ATECO and RAMO

(Figg. 2.8 and 2.9) there is also some classes prevailing consistently but anyway there is a clear sign of many more categories that suffer more than other when dealing with repaying their debts. All this data will come in handy when trying to build the best model possible in terms of performance and explainability.

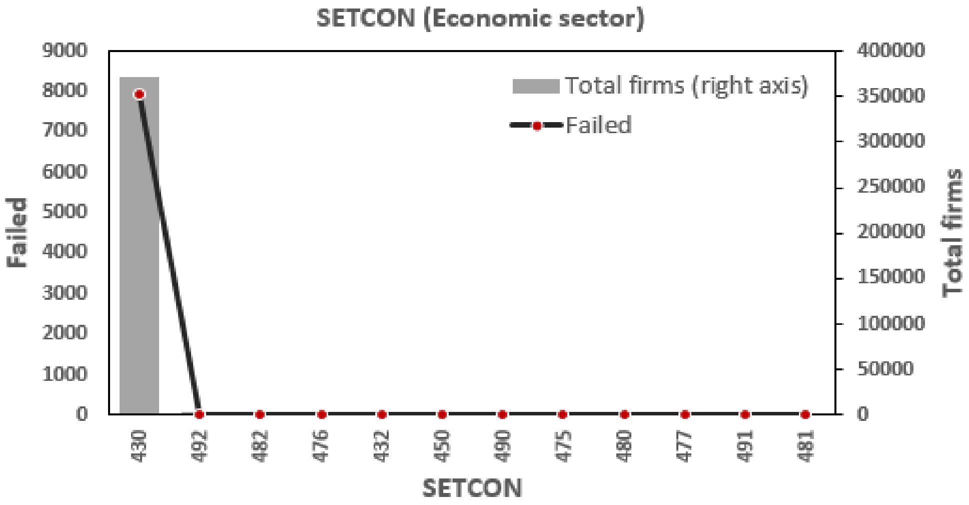


Figure 2.7: Ratio Adjusted default firms per SETCON (Economic sector) codes.

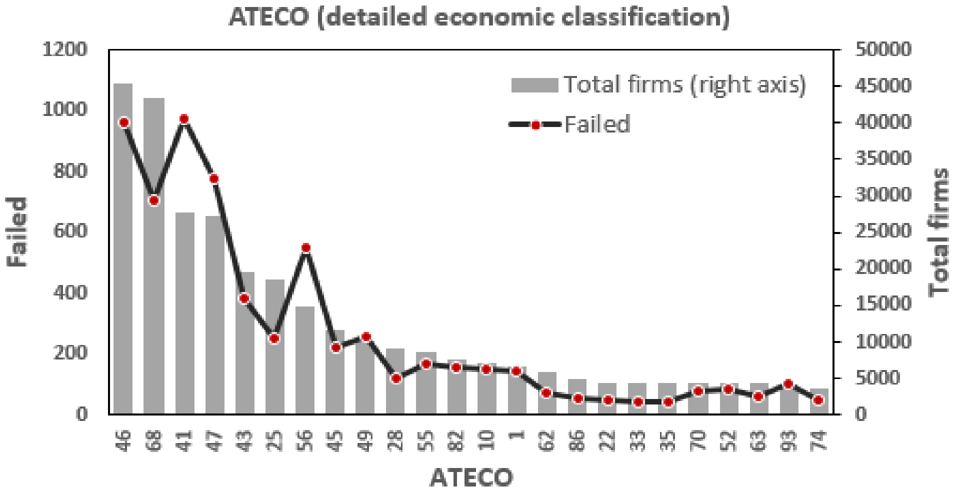


Figure 2.8: Ratio Adjusted default firms per ATECO codes.

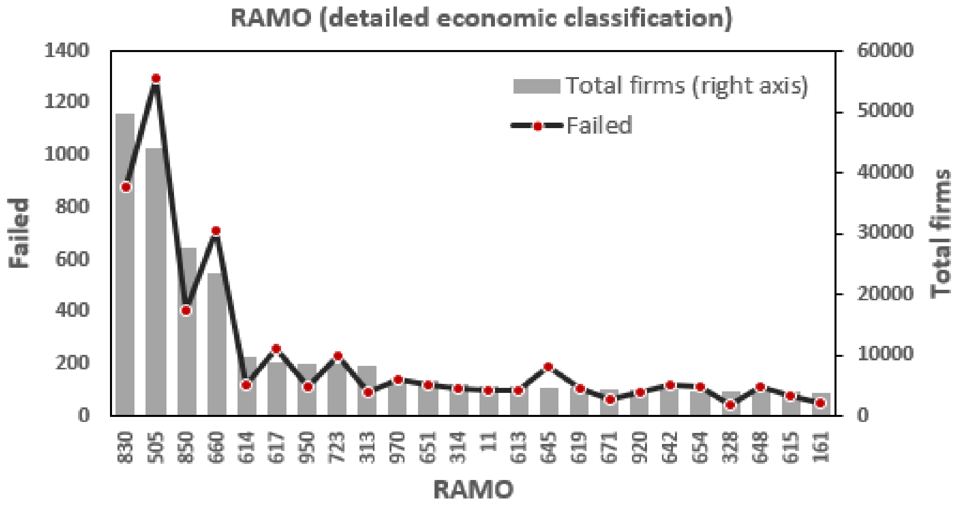


Figure 2.9: Ratio Adjusted default firms per RAMO codes.

Chapter 3

Methods and techniques

3.1 Dataset preprocessing and Feature engineering

It is really rare that an observed phenomenon is able to provide immediately the right information to be understood and modeled. Often the truth of it lies between some aspects not immediately observable that may be found only after a careful study and a deep analysis The feature engineering process consist in the all the processing steps that transform raw data, that are the result of a data collection effort, into features that can be used in machine learning algorithms and can clearly help the training process of a model to improve the performance. During the feature engineering and after a careful process of data analysis the aim of this important part of the work consists of three main steps: Feature Creation, Transformations and Feature Selection:

- **Feature Creation:** Creating features involves identifying the variables that will be most useful in the predictive model. This is a subjective process that requires human intervention and creativity. Existing features are mixed via addition, subtraction, multiplication, and ratio to create new derived features that have greater predictive power.

- **Transformation:** Transformation involves manipulating the predictor variables to improve model performance; e.g. ensuring the model is flexible in the variety of data it can ingest; finding the right balance between classes; handling the presence of missing values with the right imputer; making the model easier to understand; improving accuracy, and handling carefully categorical data paying more attention oh those with high cardinality.

- **Feature Selection:** Feature selection algorithms essentially analyze, judge, and rank various features to determine which features are irrelevant and should be removed, which features are redundant and should be removed, and which features are most useful for the model and should be prioritized.

In this chapter the main steps taken for the problem are presented.

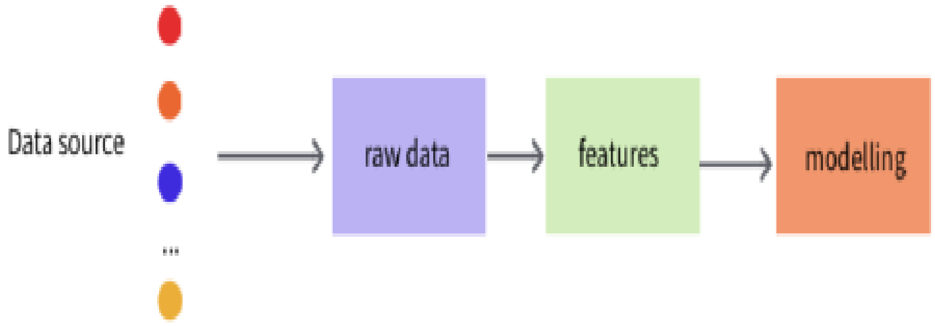


Figure 3.1: Process that leads from the source of the data to the model.

3.1.1 Missing data

In all the datasets under our study there is presence of missing data. The impact of those data differs between the credit data from the balance sheet data since in the latest their presence is much higher and my affect negatively all our process of building the best model possible. Instead for credit datasets (CCR and AnaCredit) we can assume negligible the impact of missing values.

In the Figure 3.2 we report the percentage of missing values in the Balance dataset broken down by features. We can observe that for a significant number of features we have available less than half of the potential data.

The good news is that we still have a significant share of what will prove to be the most important features available (for example: rating). Although obviously the judgment on the relevance of the less available variables is based on less robust evidence.

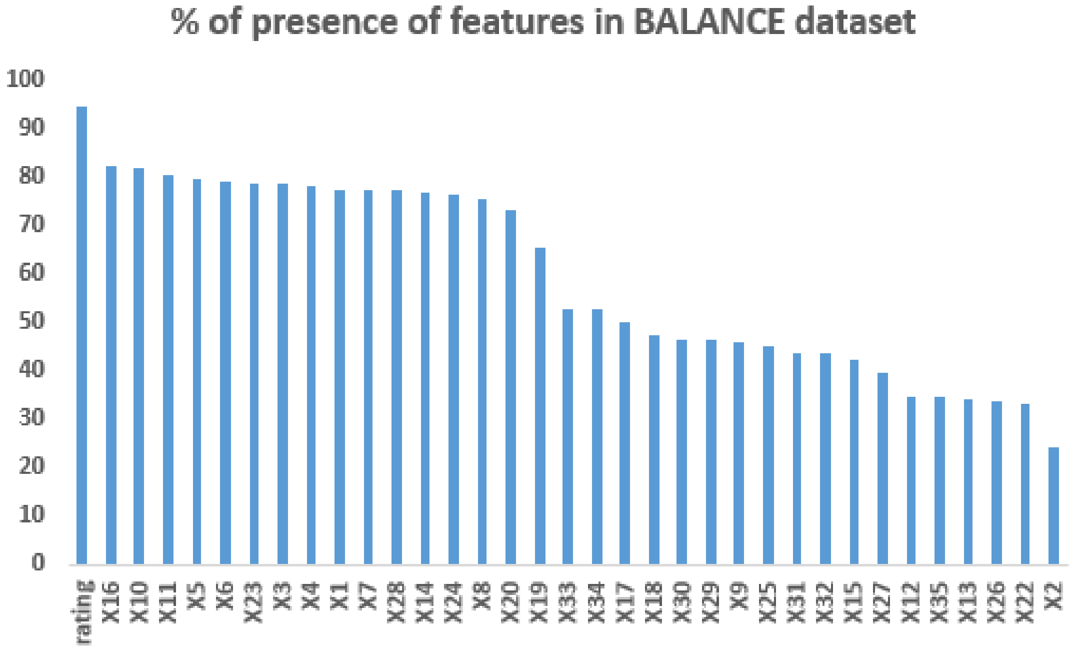


Figure 3.2: The Figure show the relevance of missing values for the Balance dataset.

In this section is going to be explored what are the possible steps to perform in order to overcome this problem and finally explain the one decided to use.

Dropping companies with missing data When a training set is large enough and there are just a few instances with missing data, one of the possible solutions may be to drop those instances. Even though it may still lead to a sound dataset and the numbers might show that the risk of the additional error is not worth the extra effort imputing the missing values, this can be very dangerous in the general case, as instances with missing data tend to be representative of particular phenomena in your data collection.

Imputing with constant values When a value, that is missing, a good imputer does not have to substitute it with a value that will be picked up as a strong signal by the model but instead you want the ML algorithm to ignore this feature as much as possible. One way to accomplish that is replacing it with a value that makes it look as nondescript as possible. In order to obtain this the easiest way is by replacing it with the average value (the mean or the median if there are strong outliers) or the most common value (the mode). While this approach is very simple, it is still much better than leaving the values as zeroes that may be interpreted as a low value very important for the classification.

Training a series of models for imputation This is by far the most complete method to reduce the error in replacing missing data and obtain the best out of it. A regressor or a classification model are trained over the entire dataset changing the target variable to the one of the features that are trying to impute. A popular and reliable choice is to use a k-NN algorithm that found its basements on the simple intuition that instances similar to the

current instance have this feature defined in a certain way. Even though this approach may seem perfect it does not scale well with the dimension of the dataset.
In the specific case of our merged dataset formed by more than three hundred thousand companies and more than one hundred columns it becomes really computationally expensive and the risk of introducing additional error make it falls in the case in which this choice was not doable.

What is done here Once all the possible steps has been taken into consideration for the dataset it has been decided to apply a mixed strategy that was both as much accurate as possible and at the same time doable in terms of complexity. The first thing was to remove both columns and rows that exceeded a certain amount of missing values (over the 70% of missing was chosen). After this choice the number of columns for the overall merged dataset decrease of about 20% and the number of companies of about 3%. The second step was to impute the missing values with the mean in order to make them as less relevant as possible while being present. And as third and last step for each column is introduced a categorical column representing if that data was missing or not in the original dataset, getting back some importance lost in the second step.

3.1.2 Handling unbalanced data

Unbalanced datasets are prevalent in a multitude of fields and sectors, and of course, this includes ours since the event of failing is usually very rare respect the entire amount of companies. The problem arise when machine learning algorithms try to identify these rare cases in rather big datasets. Due to the disparity of classes in the variables, the algorithm tends to classify those instances into the class with more predominance, the majority class, while at the same time giving the false sense of a highly accurate model. Both the inability to predict rare events, the minority class, and the misleading accuracy are bad signs that make the predictive models less reliable. This class imbalance problem between the majority and minority is frustrating and probably harmful, but not unexpected. It only becomes a real issue when this unbalance affects the performance of the algorithms or the models. If the classes are separable using the available features, then the distribution of the classes between them is not problematic. Besides, the problem is that models trained on unbalanced datasets often obtain poor results when they have to generalize. The reason lies in the fact that the algorithm receives significantly more examples from one class, making it to be biased towards that particular class. It does not learn what makes the other class "different" and fails to understand the underlying characteristics that allow to determine carefully the classes. The algorithm learns that a given class is just more common, making it "natural" having a greater tendency towards its classification. This means that the algorithm is then prone to overfitting the majority class. In this way models would score high on their loss-functions and if the evaluator does not carefully choose the right metrics to measure the performance of the model, is easy to fall in the so called Accuracy Paradox appears. In order to understand this paradox, imagine that if the incidence of category A is being found in 99% of cases, then predicting that every case is category A will have an accuracy of 99%. In practice, you do not have a good model; you have a model that assigns the same class to all the observations and does not generalize well making it instead precise just at 1%.

Once understood that this problems need to be faced, here there is the analysis of the two main ideas to overcome class imbalance if possible.

Under sampling The idea is to reduce the ratio of instances in the majority and minority classes. One way to approach this solution is to randomly select observations in the desired ratio. In this case, taking a random sample without replacement would be enough. Another way is to carry out an informed under-sampling by looking at the distribution of the data and selecting the observations to discard. In this last case, it is usually used a clustering technique or k-NN (k-nearest neighbors algorithm) to obtain a under-sampled dataset. This dataset includes observations of every natural cluster of data inside the majority class. This approach compared to random under-sampling may be better since with the latter may select all of one type of observation, and you will lose valuable information from the sample. During the re-sampling, you can try different ratios as each class does not have to contain the same number of observations. The last possible solution, was to perform an under-sampling based on a classifier. This method is called Instance Hardness Threshold (IHT). IHT method balances the dataset by eliminating data that are frequently misclassified.

Over sampling The other approach to solve imbalanced datasets is to over-sample the minority class. The simplest approach may be duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, the most popular approach consist in creating new examples that can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE. In this algorithm the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the segments joining all the k nearest neighbors of a minority class point. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Due to the enormous inequality of elements in our class of adjusted default/bankrupted companies, an oversampling algorithm will be asked to create near to 50 times more elements in the dataset causing more noise in the classification in an already noisy problem.

What have we done In our experiments reported in Chapter 4 we used the under sampling approach in line with some relevant literature. As mentioned before, in Chapter 6 we try to make something different, using instead a unbalanced training set in default prediction experiments.

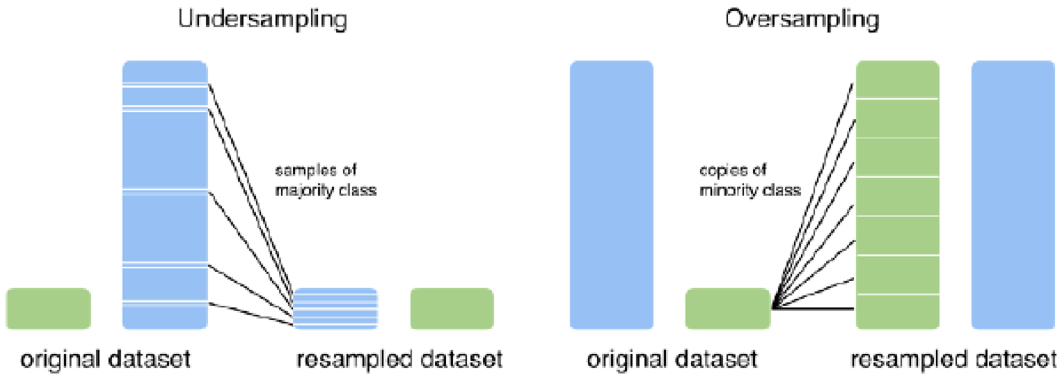


Figure 3.3: Two main approaches to handle unbalanced data: under sampling and over sampling.

3.1.3 Building up new features

While presenting the concept of feature engineering has been told that a fundamental part of it is the feature creation. Some times very representative data are hidden behind collected data and it is just needed to find it and provide it to the model. In this case, while studying statistically the three datasets it has been said that most of the data are just the absolute value of a quantity that may represent the amount of loan granted or used for example. Even though this may be really important, at the same time may not provide the full picture of a company status. For this reason, for the Central Credit Register Dataset has been decided to create one new feature and two new classes of features.

Loan dimension After some analysis has been found out that there were some ranges of granted amount of loans in which the ratio of failed companies over the total changed significantly. What has been decided empirically is that those three ranges would define a sort of dimension of the loan that may characterise the probability of default of a company. Those three ranges are:

- 0 to 250000: Here the ratio is around 2.1%;
- 250001 to 5000000: again ratio around 2.1%;
- over 5000000: ratio drop to 1.6%.

In this way there is the chance to create a new categorical variable composed by this three states.

Ratio features Another important decision around the amount of loans granted has been to determine the ratio of all the other features linked in some way to it, computing the ratio of the specific feature respect to the amount granted. This new feature has been called ratio_feature_Name and here is the list of the features processed:

- Used amount of loan;

- Past due;
- Margins;
- Amount of problematic loans;
- Amount of non-performing loans;
- Value of collateral;
- Overdraft;
- Loan expenses;
- Loan arrears.

Delta features Once defined the new ratio features, there still miss what is considered an important feature that is the difference of those ratio in the period of concern. If the used amount of a loan changes as well as the amount of non performing loans or all the other feature, it represent probably an important aspect to be considered. For this reason for the above cited features has been computed the difference in the four quarters under consideration. In this way the dataset ended up with a total of almost 500 features between original and newly computed.

3.1.4 Handling high cardinality categorical features

The above step was about finding the best possible way to handle numeric data and extract from them as much information as possible. Another matter of concern that still may contain relevant data and that, if not correctly handled may cause harm to the model, are categorical features with high cardinality. In Chapter 2 is possible to see different categorical values (such as CAP, PVAFF, SETCON, etc.) and the ratio of failed for each category. Some important elements emerged from that analysis, and starting from that it has been decided to handle those categorical features with not the classical one hot encoding way. Performing a one hot encoding of those feature would have just increased the features space with very sparse columns and probably not providing relevant information. The way chosen to perform is to replace the category value with the ratio of failed companies found, transforming this features from categorical to numeric and giving to this numeric feature a proper meaning.

3.1.5 The final size of the datasets

The significant increase in the number of columns in the overall dataset we mentioned before is due to both the additional building variables and the historical depth of the information we use.

It is worth closing this section by mentioning that the predictive experiments that will be exposed in Chapter 4 are conducted with datasets of the following dimensions:

- Merged dataset: <379,000 rows - 472 columns>

- COMB dataset: <578,878 rows - 431 columns>
- AnaCredit dataset: <379,000 rows - 73 columns>
- CCR dataset: <746,764 rows - 390 columns>
- Balance dataset: <472,280 rows - 221 columns>

3.2 Approach and techniques

3.2.1 Machine Learning techniques

As we explain in Section 1.3, the first approaches for assessing the likelihood of companies to fail were based on some fixed scores; see, for example, the work by Altman [4]. Current approaches are based on more advanced machine learning techniques. In our project we follow the literature [7] by considering a set of diverse machine learning approaches for predicting companies default.

In Table 3.1 we report all the classifiers we used. In the following we provide a brief description of each of them.

Decision Tree (DT): one of the most popular tool in decision analysis and also in Machine Learning. A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Random Forest (RF): Random forest are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes. Random decision forests correct for decision trees' habit of overfitting to their training set.

Bagging (BAG): Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging was proposed by Leo Breiman in 1994 to improve classification by combining classifications of randomly generated training sets.

AdaBoost (ADA): AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, in 2003. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost (with decision trees as the weak learners) is often referred to as the best out-of-the-box classifier.

Gradient boosting (GB): Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization

ID method	Description
NAIVE	Naive classifier based on features correlation with target
MNB	Multinomial Bayesian classifier
LOG	Logistic Regression
GB	Gradient Boosting
RF	Random Forest
DT	Decision Tree
BAG	Bagging
CAT	CatBoost
LGBM	Light Gradient Boosting
ADA	AdaBoost
COMB	Combined method based on multiple classifiers

Table 3.1: Baselines and classification algorithms. NAIVE, MNB and LOG represents our baseline in the firs set of our experiments. We describe them in Section 4.1.1.

of an arbitrary differentiable loss function. That is, algorithms that optimize a cost function over function space by iteratively choosing a function (weak hypothesis) that points in the negative gradient direction.

CatBoost(CAT): Categorical boosting [16] is a new gradient boosting algorithm that improves other standard boosting implementations. It introduces the *ordered boosting* algorithm, an alternative to gradient boosting based on permutation and an innovative algorithm for processing categorical features. Both of them are created to overcome a problem of target leakage present in currently implementation of gradient boosting algorithm.

Light gradient boosting(LGBM): Light GBM [22] is a fast, distributed, high-performance gradient boosting framework. Unlike other boosting algorithms, it splits the trees leafwise and not levelwise. LGBM runs very fast, hence the characterization *light*. On large datasets it trains faster compared to other boosting algorithms. However, leafwise splitting may lead to overfitting, which can be avoided by specifying tree-specific hyperparameters such as the maximum depth.

Except for these standard techniques, we also combined the various classifiers in the following way. After learning two versions of each classifier, one with the default parameters of the `Python scikit` implementation and one with optimal parameters (to this end, we have used an exhaustive search over specified parameter values for each classifier, using the `sklearn.model.selection.GridSearchCV`), we execute all of them (10 in total) and if at least 3 classifiers predict that a firm will default then the classifier predicts default for that firm. The number 3 was chosen after experimentation. We call this ensemble approach **COMB**.

3.2.2 Feature Selection

Whereas the construction of new meaningful features is important, equally important is the removal of features that are not correlated with the output and that may introduce noise, which may decrease the prediction performance. For this reason, feature selection can be an important step before training a model to reduce the total set of features to a subset of meaningful ones.

To choose the features to remove, we used the Boruta [24] algorithm. Boruta is a feature-selection algorithm, which aims at lowering the probability of overfitting while promoting the most important features that have an effect on the interpretability of the developed model. In a nutshell, to select what features to keep, it creates *shadow features* by permuting the input feature values among all the input elements (companies). Features that are less important important to the classifier than the shadow features are dropped.

The main difference of Boruta form other standard technique is about the choice of the threshold from which you keep a feature or not. In this algorithm features does not compete among themselves while instead the compete against a randomized version a various number of times. We can see that as a series of n independent binary experiments (important or not) following a binary distribution. In this way it is possible to determine three area of decision:

1. Area of refusal
2. Area of uncertainty
3. Area of acceptance

Technically the algorithm is designed as a wrapper around a Random Forest classification algorithm. It iteratively removes the features which are proved by to be less relevant than random probes.
The Boruta algorithm consist of these following steps:

1. Augment the dataset by adding shadow variables that are just copies of all variables.
2. Shuffle the new attributes to remove their correlations with the target value.
3. Train a random forest classifier on the extended dataset and gather the Z scores computed
4. Find the maximum Z score Z^* among the shadow attributes and then assign a hit to every attribute that scored better than that.
5. For each variable with unknown importance perform a two-sided test of equality with Z^*.
6. Set the variables which have importance significantly lower than Z^* as ‘unimportant’ and permanently remove them from the dataset.
7. Set the variables which have importance significantly higher than Z^* as ‘important’.
8. Remove all shadow attributes.
9. Repeat the procedure until the importance is assigned for all the attributes, or the algorithm has reached the previously set limit of the random forest runs.

3.3 How to measure the results?

We use a variety of evaluation measures to assess the effectiveness of our classifiers, which we briefly define. As usually, in a binary classification context, we use the standard concepts of true positive (TP), false positive (FP), true negative (TN), false negative (FN):

	Predicted Default	Predicted Not Default
Default	**TP**	**FN**
Did not default	**FP**	**TN**

- **True Positive (TP)** equivalent with “hit” (a positive successful classification);
- **True Negative (TN)** equivalent with “correct rejection” (a negative successful classification);
- **False Positive (FP)** equivalent with “false alarm” (a positive wrong classification, Type I error);
- **False Negative (FN)** equivalent with “miss” (a negative wrong classification, Type II error).

For instance, FN is the number of firms that defaulted during a particular year but the classifier predicted that they will not default.

We now define the measures that we use:

- Precision: $\mathbf{Pr} = \dfrac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$
- Recall: $\mathbf{Re} = \dfrac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$
- F1-score: $\mathbf{F1} = 2 \cdot \dfrac{\mathbf{Pr} \cdot \mathbf{Re}}{\mathbf{Pr} + \mathbf{Re}}$
- Type-I Error: $\textbf{Type-I} = \dfrac{\mathbf{FN}}{\mathbf{TP} + \mathbf{FN}}$
- Type-II Error: $\textbf{Type-II} = \dfrac{\mathbf{FP}}{\mathbf{TN} + \mathbf{FP}}$
- Balanced Accuracy: $\mathbf{BACCF1} = 2 \cdot \frac{\mathbf{TP} \cdot \mathbf{TN}}{\mathbf{TP} + \mathbf{TN}}$
- True positive rate: $\mathbf{TPR} = \dfrac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$
- True negative rate: $\mathbf{TNR} = \dfrac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}}$
- Area under the ROC curve: **AuROC**: *The ROC curve shows the TPR value at various levels of TNR, and AuROC is the area under this curve. A classifier that randomly classifies documents has an expected AuROC score of 0.5, whereas a perfect classifier has an AuROC score equal to 1*

- Matthews correlation coefficient **MCC**: *The coefficient takes into account all the 4 component of the confusion matrix (True Positive, True Negative, False Positive and False Negative) and is regarded as a balanced measure which can be used even if the classes are of very different sizes*: $\mathbf{MCC} = \frac{\mathbf{TP} * \mathbf{TN} - \mathbf{FP} * \mathbf{FN}}{\sqrt{(\mathbf{TP} + \mathbf{FP}) * (\mathbf{TP} + \mathbf{FN}) * (\mathbf{TN} + \mathbf{FP}) * (\mathbf{TN} + \mathbf{FN})}}$

Chapter 4

Experimental results

In the following part of this work, we present the main experimental results of our work.

In particular, we refer to:

- Our predictions results on bank default and bankruptcy (in this Chapter 4).
- The explanation of the forecasts obtained (Chapter 5).
- An analysis related to the prediction in our unbalanced scenario (Chapter 6).
- Some practical applications of our predictive framework (Chapter 7).

This Chapter 4 is dedicated to the description of results of default prediction; it contains two distinct sections of experiments. The first one was an important test phase of our datasets and allowed to analyze the performance of a series of ML classifiers, also in comparison with some more traditional statistical techniques. In the second section, the work was finalized, through a deep phase of data pre-processing (see Section 3.1) and the use of some additional ML algorithms, to obtain the maximum possible result in terms of predictive performance.

4.1 Early experiments

In this section we present a first set of experiments. We started our work on default prediction using some relevant classifiers in order to compare predictive performance. We used both well known statistical classifiers and the best ML classifiers; also using some simple classification methods as a baseline, which we will describe better below.

We try to predict adjusted default of Italian firms using both CCR dataset and Balance dataset, also in combination.

As a first finding, we confirm the best performance of ML methods respect to old statistical ones, in line with our reference literature (see for example [7]). In particular, the mentioned paper of Barboza and Altman represents our milestone in this first part of the work and also for this reason we use some specific performance indicators (like Type-I and Type-II errors) in order to better compare the experimental results.

Moreover, we show that the combination of credit data in addition with balance sheet data allows to improve significantly the prediction performances. The final results related to adjusted default prediction are in line with the best experimental results in [7]. It's worth mentioning that in that paper the authors try to predict bankruptcy using only a balance sheet dataset. In the second part of this chapter we will extend our experiments in this sense.

Finally we remark that, in our first set of experiments, we perform the predictions using both imbalanced and fully balanced training set highlighting significant differences in the results obtained. In our opinion this issue is extremely relevant in the context of binary prediction and is not sufficiently detailed in the literature. As we already anticipated, we will significantly extend the analysis on this topic in Chapter 6.

4.1.1 Baselines

We evaluated the techniques presented in Section 3.2.1. To assess their effectiveness, we compare them with three basic approaches. The first one is a simple multinomial Naïve Bayes (**MNB**) classifier. The second is a logistic regression (**LOG**) classifier. Finally, we created the following simple test. We first measured the correlation of each feature with the target variables (refer to Table 2.1 and Table 2.3). We found the most significant ones, (i.e., the ones that are mostly correlated with the target variable) are $C6$ (Margins) and $C5$ (Overdraft) for the CCR dataset and $X0$ (rating) and $X11$ (Margin ov revenues) for the Balance dataset.

Then we built the simple classifier that outputs *default* if at least one of the two features are nonzero and *not default* otherwise for the CCR dataset. We call this baseline **NAIVE**.

We gather the classification approaches that we use in Chapter 3 (see Table 3.1).

4.1.2 Adjusted Default prediction

We are now ready to predict whether companies will enter into an adjusted default state, as we explain in Section 1.2.3. For these first set of experiments we use only our CCR and Balance datasets.

First, we present the results for the original, unbalanced dataset. In Table 4.1 we present the results when we use only the credit dataset, whereas in Table 4.2 we present the results when we also use the balance sheet data. The first finding is that the evaluation scores are rather low. This is in accordance to all prior work, indicating the difficulty of the problem. We observe that the machine learning approaches are better than the baselines and the various algorithms trade off differently over the various evaluation measures. Random forests perform particularly well (in accordance with the findings of Barboza et al. [7]) and our combined approach (**COMB**) is able to trade off between precision and recall and give an overall good classification. Comparing Table 4.1 with Table 4.2 we see that the additional information provided by the balance sheet data helps improve the classification.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.25	0.11	0.16	0.89	0.04	0.54
MNB	0.95	0.05	0.09	0.95	0.02	0.52
LOG	0.44	0.01	0.02	0.99	0.01	0.50
GB	0.63	0.22	0.33	0.78	0.01	0.61
RF	0.61	0.21	0.31	0.79	0.01	0.60
DT	0.27	0.29	0.28	0.71	0.03	0.63
BAG	0.53	0.19	0.28	0.81	0.01	0.59
ADA	0.56	0.20	0.30	0.80	0.01	0.60
COMB	0.52	0.32	0.40	0.68	0.01	0.66

Table 4.1: Unbalanced training set; CCR data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.29	0.14	0.20	0.89	0.06	0.55
MNB	0.95	0.06	0.09	0.95	0.03	0.52
LOG	0.46	0.02	0.03	0.99	0.02	0.50
GB	0.63	0.23	0.34	0.77	0.01	0.61
RF	0.68	0.25	0.37	0.75	0.01	0.62
DT	0.28	0.32	0.30	0.68	0.04	0.64
BAG	0.59	0.21	0.31	0.79	0.01	0.60
ADA	0.61	0.26	0.36	0.74	0.01	0.63
COMB	0.55	0.36	0.43	0.64	0.01	0.67

Table 4.2: Unbalanced training set; CCR and Balance data. Higher values are better, except for Type-I and Type-II error.

In Tables 4.3 and 4.4 we present the results for the balanced dataset. There are some interesting findings here as well. First, as expected the classification accuracy improves (similarly to [7]). Second, notice that the **NAIVE** classifier performs well (expected, as feature C3 takes into account several factors of the company's behavior); however the type-II error is high. Overall, **COMB** approach remains the best performer.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.24	0.78	0.37	0.28	0.50	0.62
MNB	0.43	0.08	0.14	0.88	0.03	0.51
LOG	0.36	0.21	0.26	0.79	0.03	0.59
GB	0.23	0.67	0.34	0.33	0.10	0.78
RF	0.16	0.73	0.26	0.27	0.17	0.78
DT	0.10	0.69	0.17	0.31	0.30	0.69
BAG	0.16	0.69	0.25	0.31	0.17	0.76
ADA	0.24	0.65	0.35	0.35	0.09	0.78
COMB	0.20	0.69	0.31	0.31	0.13	0.78

Table 4.3: Balanced training set; CCR data. Higher values are better, except for Type-I and Type-II error.

	Pr	Re	F1	Type-I	Type-II	BACC
NAIVE	0.25	0.77	0.38	0.23	0.49	0.64
MNB	0.44	0.09	0.15	0.91	0.03	0.53
LOG	0.36	0.22	0.28	0.78	0.03	0.60
GB	0.19	0.78	0.30	0.22	0.15	0.81
RF	0.18	0.80	0.30	0.20	0.16	0.82
DT	0.10	0.71	0.18	0.29	0.26	0.72
BAG	0.17	0.75	0.27	0.25	0.17	0.79
ADA	0.18	0.76	0.29	0.24	0.16	0.80
COMB	0.19	0.84	0.31	0.16	0.16	0.84

Table 4.4: Balanced training set; CCR and BALANCE data. Higher values are better, except for Type-I and Type-II error.

4.1.3 Combine multiple classifiers to improve performance

Furthermore we consider an algorithm that combines several Machine Learning classifier in order to improve prediction performance. In particular we tested for ten times alternating the ML classifiers, repeating twice each of them slightly changing the parameter setting. The final result of the classification is shown in the figure 4.1 below.

At the end of classification cycle, each firm in the database can be classified in a future default status from zero to ten times. So the classification we can obtain is a function of the chosen threshold we would consider.

The overall indicator F1-score shows a maximum value for the threshold equals to three. At this threshold value we can obtain a Precision equals greater than 0.5. Increasing the threshold we can improve considerably the Precision up to almost 0.9. Of course at the price of worsening both the recall and the F1-score. Our "combined" system allows an overall gain

Combination of ML Classifiers

Threshold	Recall	Precision	F1-score
1	0.51	0.28	0.36
2	0.45	0.34	0.39
3	0.34	0.53	0.42
4	0.31	0.58	0.41
5	0.28	0.63	0.39
6	0.24	0.66	0.35
7	0.20	0.69	0.31
8	0.16	0.73	0.26
9	0.12	0.76	0.21
10	0.09	0.81	0.15

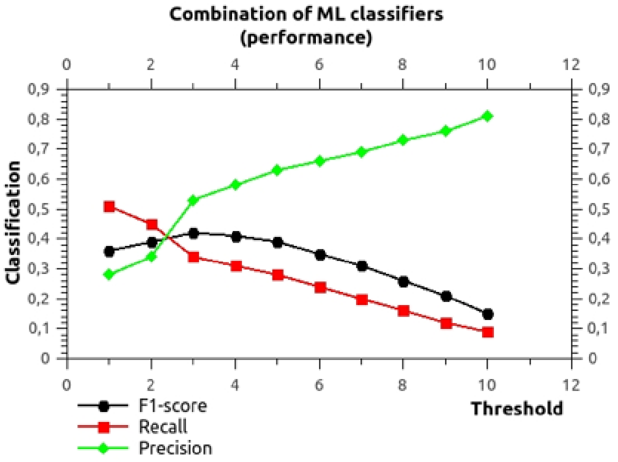


Figure 4.1: The **COMB** algorithm run for ten times alternating the basic ML classifiers. The table and the graph in this figure show that the classification we can obtain is a function of the chosen threshold we would consider. We can observe a maximum for F1-score if we choose a threshold equal to 3.

of F1-score of 5 percentage points respecting the best ML classifier tested in the previous experiments (0.43 respecting 0.37 for Random Forest classifier).

4.2 More detailed experiments

In this section we report the experimental results of a second set of experiments in which we expand both the data used and the target variables we consider. In fact, in the following experiments we use also AnaCredit data and we focus our attention on prediction of two different but related problems: corporate bankruptcy and bank default. In Section 1.2 we have analyzed in detail the links between the two conditions of vulnerability for a company.

Our starting point was the basic classification we performed in the previous section. But we enrich our analysis adding a set of new ML algorithms and performing an accurate data pre-processing.

We start by predicting whether a company will go bankrupt and next we look at the prediction of bank default.

4.2.1 Bankruptcy prediction

First, we address the problem of predicting firms' bankruptcy, which is by far the most common of the two problems addressed by past works. Our goal is to evaluate the extent to which this can be done and the value of the different types of data for deciding it. In our experiments, we use the data generated by June 2019 to predict whether the company will go bankrupt by June 2020.

In Table 4.5 we can observe the prediction results we obtain using our three credit datasets: CCR, AnaCredit and their combination Mixed. In particular, AnaCredit dataset shows a slight higher predictive capability respect to CCR dataset while the combination of the two (Mixed dataset) would seem not to provide a significant gain in performances. The best result we obtained is AuROC equal to 0.889 using **CAT** classifier.

	CCR dataset			AnaCredit dataset		
	AuROC	TPR	TNR	AuROC	TPR	TNR
LOG	0.691	0.685	0.628	0.771	0.656	0.720
LDA	0.749	0.629	0.710	0.737	0.432	0.848
DT	0.734	0.674	0.566	0.800	0.632	0.818
RF	0.854	0.764	0.767	0.845	0.728	0.812
CAT	**0.869**	**0.791**	**0.788**	**0.885**	0.745	**0.850**
ADA	0.844	0.764	0.779	0.844	**0.769**	0.842
GB	0.855	0.764	0.750	0.848	0.744	0.801
LGBM	0.837	0.744	0.776	0.842	0.728	0.783

	Mixed dataset		
	AuROC	TPR	TNR
LOG	0.714	0.657	0.692
LDA	0.794	0.722	0.799
DT	0.730	0.784	0.724
RF	0.833	0.801	0.821
CAT	**0.889**	**0.812**	0.845
ADA	0.810	0.769	0.842
GB	0.826	0.757	0.855
LGBM	0.839	0.783	0.810

Table 4.5: Performance of all our models in predicting firm bankruptcy using only our credit dataset. In particular we consider prediction performances for CCR, AnaCredit and the Mixed datasets. We can observe that in general using AnaCredit dataset allows a slight increase in prediction performance. Instead the combination of the two dataset would seem not to improve significantly the result.

	Balance dataset			MERGED dataset		
	AuROC	TPR	TNR	AuROC	TPR	TNR
LOG	0.803	0.761	0.794	0.695	0.593	0.709
LDA	0.809	0.789	0.763	0.766	0.714	0.745
DT	0.900	0.881	0.902	0.926	0.871	0.951
RF	0.937	**0.943**	0.906	0.957	0.929	0.894
CAT	0.935	0.940	**0.925**	**0.974**	**0.950**	**0.952**
ADA	0.935	0.910	0.898	0.972	0.950	0.951
GB	0.916	0.936	0.929	0.963	0.921	0.937
LGBM	**0.938**	0.926	0.906	0.971	0.943	0.950

Table 4.6: Performance of all our models in predicting firm bankruptcy using Balance dataset and Merged dataset. We can observe that using Balance dataset it is possible to obtain better performance respect to the use of only credit data. But it is clear that the use of a combination of balance-sheet data and credit data (Merged dataset) reach a significant gain in predicition performance (AuROC increase from 0.935 to 0.974 for the best classifier).

In Table 4.6 we report the experimental results of our prediction performed using Balance dataset and the combination of all the three dataset: Merged dataset.

The best results we obtained using only balance-sheet data (AuROC equal to 0.938 using **LGBM** classifier) are inline with prior work: in [7] for example the best result for bankruptcy prediction is AuROC equal to 0.93 with a Random Forest classifier. Here note that our decision is to use a balanced training set, in line with [7] and most previous works in the sector. This approach has as a result a high recall and low precision; given that in many important previous work the main goal is to achieve a high TPR and TNR, the usage of a balanced training dataset allows to achieve a good tradeoff between the two and ultimately it motivates the fact that we are focusing our attention on AuROC as a performance indicator.

In our experiments we use a balanced training set (Fix Imbalance Method: RandomUnderSampler) and a 10-fold cross validation.

Next, we measure the improvement that we obtain by using also the CCR and AnaCredit dataset in combination with Balance (i.e. Merged dataset); the results appear in the right side of Table 4.6.

Observe that the performance improves significantly compared to the use of the balance-sheet data alone. The best performance is achieved with **CAT**, arriving to rates of close to 0.95 for both TPR and TNR, and AuROC of 0.974. According to our knowledge, the result obtained is one of the highest even compared to the reference literature of this field of research.

These results represent a significative confirmation of the importance of combining both sets of data (balance sheet data and credit data), as they capture different dimensions of a company's health.

Moreover, an interesting result concerns the fact that although the combination of the two datasets leads to better results, the data related to the financial statements (balance sheet data) seem to be more relevant in order to predict bankruptcy. In fact, using only credit data, the performances are significantly lower. Thanks to the use of the Boruta Algorithm for feature selection we reduce the dataset to a total of just 20 significant attributes. As we can see later this will be very important for the explanation of the results.

4.2.2 Adjusted Default prediction

Recall from Section 1.2, that the adjusted default is a status that often precedes bankruptcy and a sign that a future bankruptcy is eminent and as such, there is a high interest from the banking system to predict it. In this section we follow the structure of the previous one and we evaluate the possibility to predict adjuste default, first with our credit datasets: CCR, AnaCredit and Mixed, then using Balance dataset and finally with the combination of all our primary datasets, i.e. Merged dataset.

We start by showing the results obtained using our credit datasets (see Table 4.7). In this case we observe that CCR dataset and AnaCredit dataset reach similar prediction results for the best classifiers, but the combination of two (Mixed dataset) allows a significant gain in performance (AuROC equal to 0.912 from 0.885 obtained using only AnaCredit dataset and considering the best classifier).

	CCR dataset			AnaCredit dataset		
	AuROC	TPR	TNR	AuROC	TPR	TNR
LOG	0.756	0.752	0.632	0.741	0.822	0.532
LDA	0.644	0.522	0.877	0.753	0.557	**0.870**
DT	0.855	0.700	0.866	0.858	0.791	0.778
RF	0.866	0.747	0.830	0.873	0.788	0.816
CAT	**0.881**	**0.778**	0.830	**0.885**	**0.810**	0.805
ADA	0.868	0.740	0.855	0.875	0.767	0.828
GB	0.877	0.771	**0.981**	0.880	0.805	0.814
LGBM	0.880	0.775	0.825	0.882	0.802	0.806

	Mixed dataset		
	AuROC	TPR	TNR
LOG	0.750	0.690	0.690
LDA	0.752	0.703	0.683
DT	0.854	0.708	0.815
RF	0.889	0.755	0.820
CAT	**0.912**	**0.805**	**0.890**
ADA	0.894	0.728	0.859
GB	0.889	0.773	0.817
LGBM	0.892	0.773	0.815

Table 4.7: Performance of all our models in predicting firms adjusted default, using only our credit datasets. In particular we consider prediction performances for CCR, AnaCredit and Mixed datasets. We can observe that in general CCR dataset and AnaCredit dataset obtain similar results in prediction performance. Instead the combination of the two dataset allows a significant improvement in the results (AuROC 0.912 from 0.885 for the best classifier).

	Balance dataset			MERGED dataset		
	AuROC	TPR	TNR	AuROC	TPR	TNR
LOG	0.700	0.504	0.788	0.722	0.679	0.756
LDA	0.683	0.626	0.678	0.762	0.720	0.722
DT	0.791	0.689	0.822	0.886	0.795	0.847
RF	0.815	0.766	**0.834**	0.930	0.874	**0.871**
CAT	**0.843**	0.790	0.788	**0.953**	0.890	0.885
ADA	0.824	0.786	0.788	0.941	0.865	0.870
GB	0.826	0.780	0.783	0.943	**0.894**	0.861
LGBM	0.834	**0.791**	0.781	0.941	0.876	0.863

Table 4.8: Performance of all our models in predicting firms adjusted default using Balance dataset and Merged dataset. We can observe that using balance sheet data we obtain lower performance respect to the use of credit data. But it is interesting to note that the use of a combination of balance-sheet data and credit data (Merged dataset) reach a significant gain in prediction performance (AuROC equal to 0.953 for the best classifier, while we obtained 0.912 and 0.843 using Mixed dataset and Balance dataset respectively).

In Table 4.8 we use Balance dataset. In this case, for the best classifier (**CAT**) AuROC is equal to 0.843, a value that is significantly lower respect to the results we obtained using credit datasets. In this scenario our conjecture is that credit data are crucial in order to perform a better prediction for adjusted default. But we can observe that the use of balance sheet data in addition to credit data can significantly improve our ability in prediction, also in the case of adjusted default prediction. In fact, in the right part of Table 4.8 the results related to the Merged dataset show an AuROC of 0.953 with an important improvement respect to the use of only credit data. Moreover, as we will see later, we can gain also in explainability of the results. In a similar but symmetrical way with respect to the previous case, we can conclude that the balance sheet data can help us to improve bank default prediction but credit data remain essential for this issue.

It should be noted that, also in this case, the use of Boruta Algorithm leads to a drastic reduction of the relevant features that remain about 20 (see Figure 5.1). In our opinion, the message from both prediction problems is clear. Company failure prediction is a hard problem, and relying on the public balance sheet data has its limits. Results can improve when we take into account the past behavior of the companies with respect to loan repayment.

4.3 Past credit information is critical in predicting default, but too old is no longer useful

In this section we are interested in understanding how relevant older information is in default prediction. And therefore how important it is to have available very long time series when we use ML algorithms.

First of all, we tried to predict adjusted default using only credit data from CCR dataset. In particular, we have tried to predict companies that will default in June 2020 (and that are in a good banking situation in June 2019), using company credit data starting in June 2014. The results are reported in Fig. 4.2 in which, for simplicity, we have taken into account only AuROC as performance indicator.

We can observe that using data relating to the most recent quarters (i.e. in our particular case using only the data of June 2019 to predict a possible default in June 2020) we obtain a value for the AuROC that is already quite high but which is not yet the maximum obtainable.

As we can see from the Fig. 4.2, however, by increasing the information, going back in time, the predictions improve. However, this improvement only appears up to about 4-5 quarters backwards. From that point on, using additional credit information no longer improves the results in a significant manner, as can be seen by looking at the right side of the mentioned chart that report the results obtained using up to five years of past credit data. This experimental result appears interesting to us although not really easy to explain. A possible conjecture is that too old credit information relating to the individual company is not useful for predicting the firm default as the negative signals, when they materialize, fulfill their negative effect in a relatively short period of time. We think that the phenomenon seems worthy of being investigated also from an economic point of view. But for our scope, we can certainly affirm that the ML models we currently use cannot take into account any information useful for the prediction of bank default that is too old in time.

We conclude anticipating that from the analysis relating to explainability (Chapter 5) it clearly emerges that the most recent information is largely predominant in the forecast.

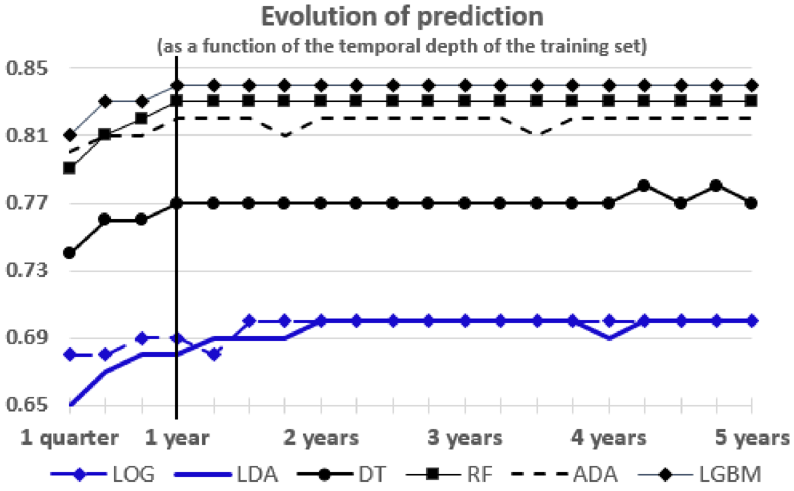


Figure 4.2: Adjusted default prediction using quarterly credit data. Evolution of predictive performance of all our models in predicting adjusted default as the depth of the dataset used for predictions varies. We can observe that the best performances are obtained using about 4 quarters of data and do not significantly improve as the data used increases.

In the following Figure 4.3 we report the similar results obtained by performing bankruptcy prediction experiments, using only balance sheet data. Also in this case we use a time series of five years, but of annual data. We get a confirmation that using a longer time series does not add significant performance improvements, also in the case of balance sheet information. Although in this case the message seems more uncertain as for some important classifiers (such as RF and LGBM) there are small predictive improvements by adding older data to the training set used in prediction.

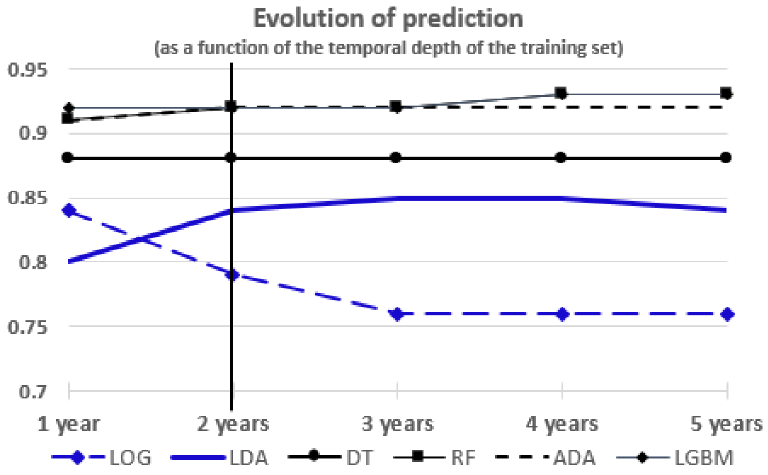


Figure 4.3: Bankruptcy prediction using annual balance sheet data. Evolution of predictive performance of all our models in predicting bankruptcy as the depth of the dataset used for predictions varies. Also in this case we can conjecture that the use of a more long time series of balance sheet data do not significantly improve the prediction performance.

4.3.1 Performance prediction with a smaller sample of data

So far we have used a huge amount of data for default predictions. However, the use of such large datasets is not always possible for all stakeholders interested in predicting default. In principle, a National Supervisory Authority or a Government may have similar datasets available, but a private bank may only have a portion of this data available.

In this section we try to verify how much the size of the dataset impacts on predictive performance. We compared adjusted default predictions obtained using only a subset of our complete dataset (Merged). In Table 4.9 we can see the results obtained. In particular, we observe that using only 10% of the data (i.e. information relating to about 40,000 companies) the performance remains substantially unchanged. If, on the other hand, we use 1% of the dataset (about 4,000 companies) we find a slight reduction in the predictive capacity (the AuROC drops to 0.89 from 0.95). Finally, the performance is drastically reduced (AuROC equal to 0.75) if we use a dataset that contains information relating to only 1,000 companies.

Therefore, our conjecture is that it is possible to obtain good predictive ability even using only a subset of our datasets. In particular, the use of data from around 4,000 companies still allows good performance.

In Table 4.10 we show a breakdown of the banks operating in Italy divided by number of customers belonging to the category of companies. We can observe that a small share (2.6%) have information of more than 40,000 companies while about 20% can have a dataset of more than 4,000 companies. Finally, about 40% of the banks have information on more than 1,000 firms available. This subdivision only examines banks without considering membership of banking groups which could allow individual banks to have even wider firms datasets available.

	MERGED dataset			10% of MERGED			1% of MERGED			Only 1,000 firms		
	AuROC	TPR	TNR	AuROC	TPR	TNR	AuROC	TPR	TNR	AuROC	TPR	TNR
LOG	0.72	0.68	0.76	0.69	0.64	0.71	0.65	0.61	0.67	0.59	0.55	0.62
LDA	0.76	0.72	0.72	0.74	0.70	0.71	0.71	0.68	0.69	0.61	0.59	0.62
DT	0.89	0.80	0.85	0.88	0.78	0.84	0.81	0.75	0.80	0.67	0.65	0.66
RF	0.93	0.87	0.87	0.93	0.87	0.87	0.87	0.83	0.84	0.73	0.71	0.74
CAT	**0.95**	**0.89**	**0.89**	**0.94**	0.88	**0.89**	**0.89**	0.85	**0.84**	**0.75**	**0.75**	**0.76**
ADA	0.94	0.87	0.87	0.93	0.86	0.87	0.87	0.85	0.84	0.73	0.74	0.73
GB	0.94	0.89	0.86	0.93	**0.89**	0.86	0.88	**0.86**	0.82	0.73	0.72	0.75
LGBM	0.94	0.88	0.86	0.94	0.87	0.86	0.88	0.84	0.83	0.74	0.74	0.73

Table 4.9: Evolution of predictive performance in predicting adjusted default as we reduce the database used for predictions. In particular in the table we compare the results obtained using the total dataset (Merged) with the prediction results we obtain using a 10% of the dataset (i.e information related to about 40,000 firms), 1% of the dataset (about 4,000 firms) and a small dataset composed of only 1,000 firms.

	Italian banks for number of *Non financial corporations* debtors				
	> 40,000 firms	4,000 - 40,000 firms	1,000-4,000 firms	< 1,000 firms	< 20 firms
Number of banks	10	69	87	67	157
% of total	2.6	17.7	22.3	17.2	40.3

Table 4.10: Breakdown of Italian banks by number of debtor companies (as of December 2021). About 20% of banks have over 4,000 debtors belonging to the *Non financial corporations* (NFC) category in their customer portfolios. These banks account for over 90% of the credit to NFC in Italy.

4.4 Some details about the experiments

Here, we give some little information to the practical experiments we performed, referring to the code for all the technical details in the use of the programs.

- We use generally the tool Scikit [40] for the early experiments.
- Moreover, we use PyCaret and the Colab environment with reference to the second part of the experiments. In particular, PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It is an end-to-end machine learning and model management tool that exponentially speeds up the experiment cycle and makes you more productive.
- In general, in order to tune the hyperparameters, we split the datasets into training, validation and test sets. Specifically, we typically performed a random stratified split of the full training data into train set (80%) and validation set (20%)

4.5 Where are the neural networks?

We would like to close this chapter dedicated to our experimental results with a more general consideration. Indeed, after all this sets of results, metrics and analysis, some of the readers may have a common question: where are the neural networks? Deep neural networks have demonstrated great success across various domains that have long been considered a challenge. Several highly well performing architectures exist for these problems that succeed in encoding raw data in real-world application efficiently into meaningful abstraction or representation.

In NLP, deep neural networks models are now state-of-the-art, outperforming older and conventional machine learning algorithm. Models such as BERT, RoBERTA, XLNet are the new gold standard and used every day such as in Google that has deployed BERT in its search engine in 2019, making one of the most largest update to its search engine in the past few years and it wouldn't take long now that these solutions will become commonplace in commercial products. It is possible to see practically that deep neural networks for computer vision are currently affecting our day life through cars with self driving capabilities or more trivially in our smartphone with filters on social networks and other applications thanks to generative adversarial networks (GANs) like StyleGAN. Worth of notice may be also AlexNet, winning the ImageNet challenge in 2012 or ResNet in 2015 achieving superhuman accuracy. It is clear from these initial part that deep neural networks have taken over the domain of unstructured data.

Another story must be told for tabular data or better called structured datata. These data consist of a set of samples (rows) with the same set of features (columns), they usually are highly organized in a tabular structure to allow efficient operations on the table columns such as search and joins such as in SQL databases. It is the most common data type in real-world applications such as ours in the financial field. Many challenges arise when applying deep neural networks to tabular data:

- Usually there are heterogeneous, the features constructed from tables come from various unrelated sources, each with their own units and associated numerical scaling issues.
- Features are often correlated and usually just a small subset of features are responsible in the prediction.
- The common presence of highly unbalanced datasets and the relative difficult to enhance them to overcome this problem. A lot has been analysed in previous chapters; in addition we dedicate to this issue the entire Chapter 6.
- Data are usually sparse and that lead to features in a high-dimensional space that is generally not dense and continuous, making it difficult to exploit for a typical deep neural network.

Another crucial point is that for tabular data related problems, prediction as already seen is not the main goal, but providing the right explainability to a prediction is just as, if not more, important and neural networks are mostly seen as black-box models.

Recently many scholars had focused on trying to extend the use of deep neural networks to tabular data, emulating anyway the best aspects of tree-based algorithm. Some of the most promising works are: TabNet, NODE (Neural Oblivius Decision Ensembles), DNF-net. The results of their researches seems to be at least comparable to one obtained with tree-based models. Anyway in their work "Tabular Data: Deep learning is Not All You Need" [41] shows that in general modern tree based algorithm like XGBoost outperforms these deep neural network in more that one tests, including datasets used in the papers that proposed the models. They also outline that XGBoost for example requires much less tuning and it is much more faster. In the end is still possible to say that the state-of-the-art for tabular data are tree based machine learning models and that deep learning methods still needs some work to do.

4.6 An evaluation of the results

It is not an easy task to evaluate the quality of the performances obtained. The issue of predicting default is complex. The reference literature is endless and the approaches used, also with reference to the measurement of the results, are very numerous.

Let's try to make a small comparison, without any pretense of being exhaustive, in the following way. As already mentioned, we used Altman's 2017 paper (see [7]) as our reference work, which we consider our benchmark. We recall that in Altman's paper the authors obtain a 0.929 in AuROC as best result, using a Random Forest classifier and a balanced training set (see Table 4.11). Based on this previous work, we oriented our predictive experiments using a balanced training set and trying to obtain maximum performance by working on the pre-processing of the datasets and the selection of the techniques used. Our goal was to improve the final result of our prediction by trying to explore our upper limit. We have collected in table 4.11 a selection of papers considered by us. Among these we considered the most relevant ones citing Altman's work and two works using data from the Italian CCR. A common trait of the works we select in the following table is to provide results using AuROC as a performance indicator. Which makes the results better comparable with our work. In addition we include in the comparison also a 2006 paper (see [45]) that we will use as our reference point in the analysis related to the prediction in unbalanced scenario, in Chapter 6.

The comparison of our results with the papers mentioned in Table 4.11 confirms that the performance obtained in the prediction of business bankruptcy (AuROC equals to 0.974) are of absolute importance. It is instead more difficult to compare the forecasts relating to bank default where there is still no significant reference literature. Our results (AuROC equals to 0.953) also in this case appear to be of absolute importance; for example, they far outclass those obtained in [34], in which the authors try to predict a companies' bank default status slightly different from ours adjusted default definition, but still comparable.
As we anticipated, our conjecture is that it is difficult to get much further in performance using the type of data we have available. The issue of using other data in this task (which we do not deal with in this work) seems to us to be of absolute relevance. For example, the possible use in prediction of companies default of unstructured data obtainable from the Web.

Authors	**Title**	**Year**	**Dset size (rows)**	**Main ML techniques**	**AuROC**	**Quotes**	**Train set**
E. Altman et al. [7]	Machine learning models and bankruptcy prediction	2017	13,300	RF/BAG/ GB/NN	**0.929**	464	balanced
Wo-Chiang Lee [28]	Genetic Programming Decision Trees for Bankruptcy Prediction	2006	130	DT/GP	**0.899**	28	balanced
Moscatelli et. al [34]	Corporate default forecasting with machine learning	2020	260,000	RF/GB	**0.846**	54	balanced
Tuong Le et al. [26]	Oversampling Techniques for Bankruptcy Predict. Novel Features from a Transaction Dset	2018	120,000	DT/RF	**0.844**	60	balanced
S.Daskalaki et al. [45]	Evaluation of classifiers for an uneven class distribution problem	2006	n.a.	DT/SVM NN	**0.9**	195	balanced/ unbalanced
Monica Andini et al. [5]	Machine learning in the service of policy:targeting the case of public credit guarantee	2019	4000	DT/RF	**0.8**	8	balanced
H. Kvammea et al. [25]	Predicting mortgage default using convolutional neural networks	2018	21000	NN	**0.92**	111	balanced

Table 4.11: A selection of reference papers in our context. For each work, the techniques used, the size of the dataset used and the predictive performance of the default expressed in AuROC are highlighted.

Chapter 5

Explainability

This chapter is probably the most important one especially for the field of application of our models. The aim here is to provide in this thesis what is known as XAI (eXplainable Artificial Intelligence). When it comes to explainability the most of the researches are usually focused on explaining decisions to a final human observer, and it should be easy to say that looking at how humans explain to each other can be an important starting point to build explanation in artificial intelligence. In this chapter we will start providing a definition of what explainable means according to literature to end up with the definition and some examples of the approach decided to be used in this work, that is model agnostic explainability.

5.1 What explainability means

According to many scholars such as Tim Miller [32], explainability is more than one single entity or concept, instead it is a union of different concepts that may seem to be interchangeable and they have been studied for centuries. What can be said is that explainability can be expressed as the degree to which a human can understand the cause of a decision. The more a machine learning model is interpretable, the easier it is for the model's user to comprehend why certain decisions or predictions have been made. A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the another model. Why explainability is important and why it is needed in this problem has already been discussed, now it is the moment to determine how machine learning explainability can be classified and measured. What can be expected form explainable algorithms may be summarized into three different type of explanations:

1. Global model explainability: this type of requirement is to answer the question to which parts of the model affect prediction the most. It is a general overview of the model explainability.

2. Local explainability for a single prediction: This type if explanation is fundamental in problems like ours since it is needed to provide the reasons why a single decision about lending or not a loan.

3. Local explainability for a group of predictions: such as the previous points it is very important especially if someone want to study a certain phenomena such as which are the feature in a specific area that influence the outcome of the prediction in that part of the world. There may be different explanation for different zones or different type of company for example.

It is important to outline that explainability may be intrinsic or post-hoc [33]. The first type refers to machine learning models that are considered interpretable by their means, due to their simple algorithm and structure, some example may be short decision trees or simple linear models. The second one instead refers to the interpretation of the model after its training and may be applied also to intrinsically explainable models and it is more commonly know as model agnostic explainability.

5.1.1 Model agnostic explainability

Dividing the explanations from the machine learning model brings some important advantages on many levels. The first one over model-specific explainability is the model agnostic flexibility. Data scientist and Machine learning developers are free to use any machine learning model and then apply the explainablity method on any model. Since typically when trying to solve a problem many types of machine learning models are evaluated, when it comes to compare models explainability, it is easier to work with model-agnostic explanations, because the same method can be used for any type of model. This approach shows some more important aspects such as:

1. Model flexibility: The explainablity method works with any machine learning model.
2. Explanation flexibility: there are no limits to a certain form of explanation, it is a formula or a chart.

Essentially what model agnostic algorithms for explainability do is just adding another layer over the classification model that helps humans to understand what is going on. Despite the huge plethora of model agnostic methods it has been decided to focus and use to the latest studied that is based on the game theoretically optimal Shapley Values.

5.2 The need of explainability

For various applications, including financial ones, just predicting the probability of failure does not suffice and it is important to know:

1. which input features lead a company to be characterized as safe or risky (local explainability);
2. which are the input features with most predicting value overall (global explainability).

There are two approaches to provide explanations. The first one is to use an *interpretable* model, such as logistic regression or a (shallow) decision tree. Such models are fairly intuitive and by observing them the user can understand the classification results. However, it limits the class of classification functions that can be learned, leading to sub-optimal accuracy.

The second approach is to use an interpretability approach, which uses the classification as a black box. Such models, consider subsets of the features or create new input data by distorting the real ones, and they evaluate the classification accuracy on these modified problem instances, to assess the importance of the features.

For our problem we used the TreeSHAP approach [15], a modern method to provide interpretability to tree-based models. It is based on the SHAP method by Lundberg and Lee [30], and maps the problem of estimating the importance of the input features to the game-theoretic problem of distributing gains to agents who work together in a coalition. It is an approach that has some mathematical properties (see [30]) and which works well in practice in diverse fields.

Focusing just on how well a machine learning model performs is not enough. Understanding the reasons why a model makes a certain prediction is very important especially in financial applications. The user of the model needs to trust the decision proposed and must be able to explain it to the final customer.

The first criteria to analyse in the choice of the right tool is if we want to find the intepretability of the model by restricting its complexity or analyzing it after training. The second one instead is about to choose a model specific tool, like for example the interpretability power of regression weights in a linear model or a model agnostic one that can be used whatever model is chosen. The last methods does not have information about the model specifics but rely just on the input and the output. In our work, considering the already known difficulties of the problem, we decided the second chance for both the issues. We don't want to giving up on the complexity of the model while at the same time we want to be able to change it keeping the possibility to explain the results.

One way to see the problem of explainability is to imagine each feature as a player in a game and the prediction as a payout. This a classic game theory approach. We want to attribute to each feature the proportional value representing its contribution to the prediction for each given instance. This type of tools are called additive feature attribution methods and their model is a linear function of binary variables:

$$f(x') = \phi_0(f) + \sum_{M}^{i=1} \phi_i x'_i \tag{5.1}$$

where $z' \in 0, 1^M$ represent a feature, $\phi_i \in R$ is the attribution value and M is the number of features.

A method to be considered a good feature attribution method must satisfy three specific properties:

1. Local accuracy: when approximating a model f for a specific input x, the explanation's attribution values $\phi_i(f, x)$ should sum up to the output $f(x)$.

2. Consistency: in presence of changes in the model, if the contribution of a feature increase

or stays equal, its input's attribution should not decrease.

3. Missingness: features with no effect on the function $f(x)$ must have no assigned impact.

For tree-ensamble methods like gradient boosting machine and random forest, for each input feature is possible to determine an importance value. These values are shown to be inconsistent and thus don't provide clear and definite explanation over the model's prediction. Here we explain how a recent approach called treeExplainer is currently the best for this purpose.

TreeSHAP TreeSHAP from Explainable AI for Trees by Lundberg et al. (2020) [15] is a modern method to provide interpretability to tree-based models. It is based on the Shap method by Lundberg and Lee (2016)[30] that rely on the Shapley values from the game theory. Shap (SHapley Additive exPlanation) has been proved to be consistent with human intuition and are very useful due to their ability to provide insights on how each feature plays a role in a single prediction. In their research they also prove how Shap is the only possible method in the broad class of additive feature attribution methods to satisfy the properties of *local accuracy*, *consistency*, *missingness*.

Shapley values are computed by introducing each feature, one at a time, into a conditional expectation function of the model's output. The change produced is attributed at each step to the feature that was introduced. Then the algorithm average over all possible feature orderings.

Even if theoretically good for the purpose Shapley values are hard to compute. With threeShap we have the first polynomial algorithm based on trees that is able to compute optimal explanations based on game theory that directly measure local feature interaction effects and allows to understand global model structure based on those local effects.

As we mentioned earlier, being able to provide sound explanations about why a classifier has a particular output is often a crucial requirement for financial applications. In this section, we analyze the input features and we evaluate their importance through their Shap value.

5.3 Bankruptcy prediction

We analyze Bankruptcy predictions that we obtained both using the Balance dataset and Merged dataset. In both cases, we consider our best performing model (**CAT**), but using other classifiers we obtain similar results. From the explanatory overview we can observe, in Figure 5.1, the most important features that determine the forecast based on Balance dataset: the last available year's *rating* appears to be the most relevant one. The other balance sheet indicators that become significant are $X1(Revenues)$, $X8(ROE)$ and $X16(Liquidity)$.

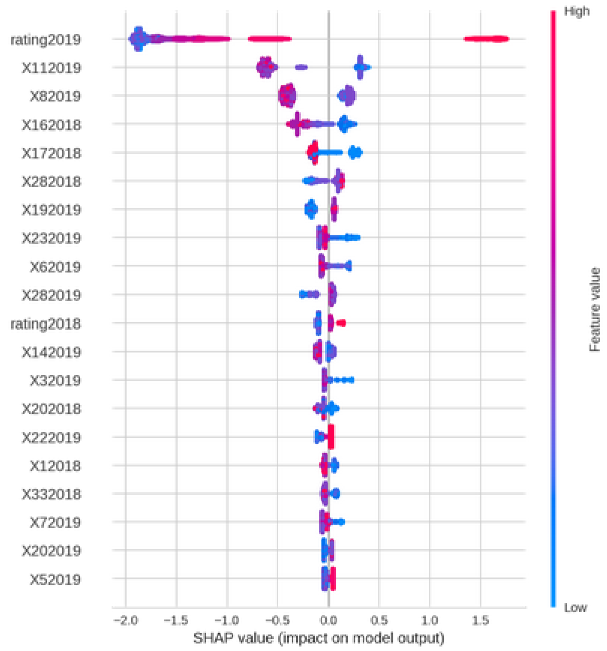


Figure 5.1: Visual depiction of the Shap values of the most important features for the **CAT** classifier for bankruptcy predictions, using the Balance dataset. Each row corresponds to one of the features with the highest Shap values. The color gradation indicates the feature value. The x-axis corresponds to the Shap value. The most powerful explanatory predictor is *rating* that is an indicator calculated on the basis of other fundamental financial indicators and represents an evaluation on the state of health of the company provided by a specialized company (Cerved).

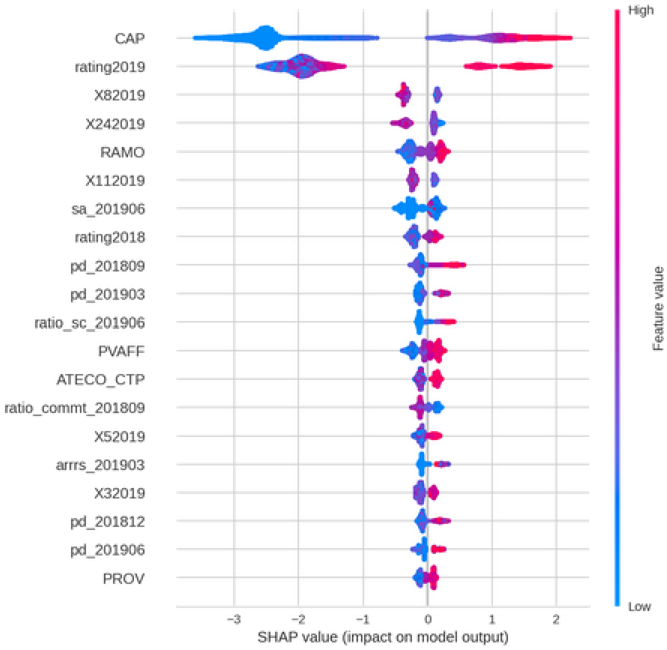


Figure 5.2: Visual depiction of the Shap values of the most important features for the **CAT** classifier for bankruptcy predictions, using Merged dataset. Each row corresponds to one of the features with the highest Shap values. The color gradation indicates the feature value. The x-axis corresponds to the Shap value. In this case we can observe a significant effect on the explainability by the detailed geographical localization (CAP) and, also in this case, a significant impact of rating.

In Fig. 5.2 we can observe the main factors that can explain the predictions obtained using our Merged dataset. The picture shows a significant impact of the detailed geographical localization (CAP), that is one of the categorical features that we consider in the Merged dataset. We provide a particular pre-processing to these categorical features that we describe in the following paragraph. We found that categorical features don't give a relevant contribute regarding the prediction performances but some of them (CAP, for example) can give an important contribute to explainability. In our opinion, this experimental evidence suggest the conjecture that prediction performances and capability of explain the results are two goals which are not always closely related.

5.3.1 Handling of categorical features

In our experiments with Merged dataset we perform a pre-processing of the categorical variables in order to better exploit their informative value. Our data manipulation consists in associating to each value of the categorical variable the failure rate relative to that variable's value. We found that this data pre-processing is not reflected in a significant increase in predictive performance but instead has significant effects on the explainability of the results. However, this happens in particular only for some categorical variables. In particular, the most evident impact on the explainability is related to the detailed geographical location (CAP code). The explanation we give to this result is connected with the fact that we found a considerable variability in the rate of defaulted firms as the geographical location varies (see Fig. 5.3).

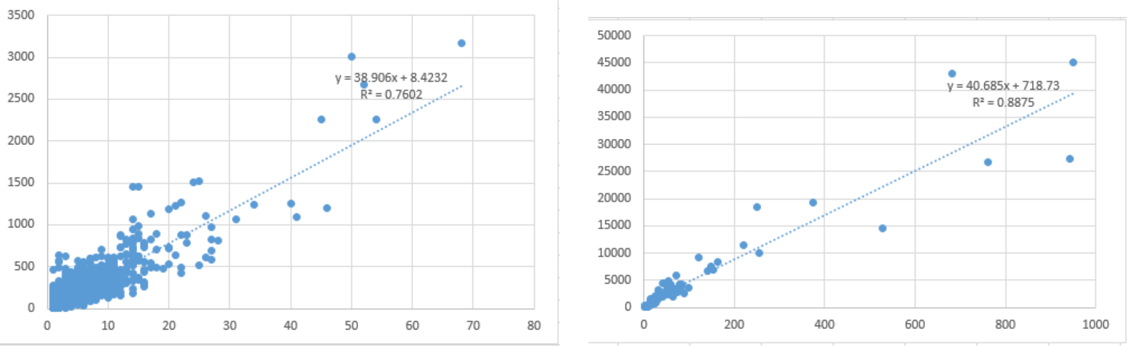


Figure 5.3: Scatter plot that shows the relation between total firms(y-axis) and failed (x-axis) for CAP variable (on the left) and $ATECO\_CTP$ (on the right). We can observe a significant difference in dispersion between the two categorical variables. When we consider detailed geographical classification we found a larger difference in rate of bankruptcy firms between the different CAP codes. This particular characteristic of our data could represents the reason why CAP code takes on such an important explanatory power.

5.3.2 Explainability of single prediction

Shap values allows us to see also in detail the effect of each attribute on each of the individual prediction. In Figure 5.4 we can see an example in which a firm is predicted that will go bankrupt (positive example). We observe that this positive prediction can be mainly attributed to two categorical features (CAP and RAMO) and, to a lesser extent, also to the rating in the last year (2019) and some others balance sheet indicators.

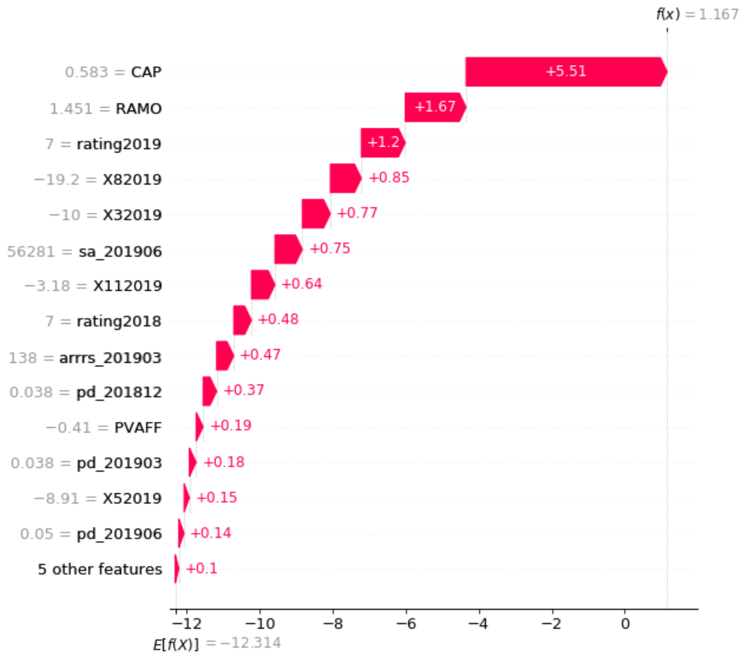


Figure 5.4: The highest SHAP values of the features for a single positive (bankrupt) prediction by the **CAT** classifier. Each slice corresponds to a single feature, the length of the slice indicates the SHAP value of the feature, and the color indicates whether the features leads to a positive classification (red) or negative classification (blue). In this example, we can observe the significant impact of two categorical features (CAP and $RAMO$) plus the effect of a bad rating classification. To be noted that in this particular case, the SHAP algorithm did not identify any factors of opposite sign (blue lines).

Instead, in Figure 5.5 we can see an example of how the prediction of an healthy organization is primarily driven by the detailed geographical localization plus a couple of balance sheet indicators ($X24$ and $X8$) and a good value for the probability of default. These factors explain the negative prediction (no bankrupt) which occurs despite the negative rating classification, which would instead push the prediction in opposite direction.

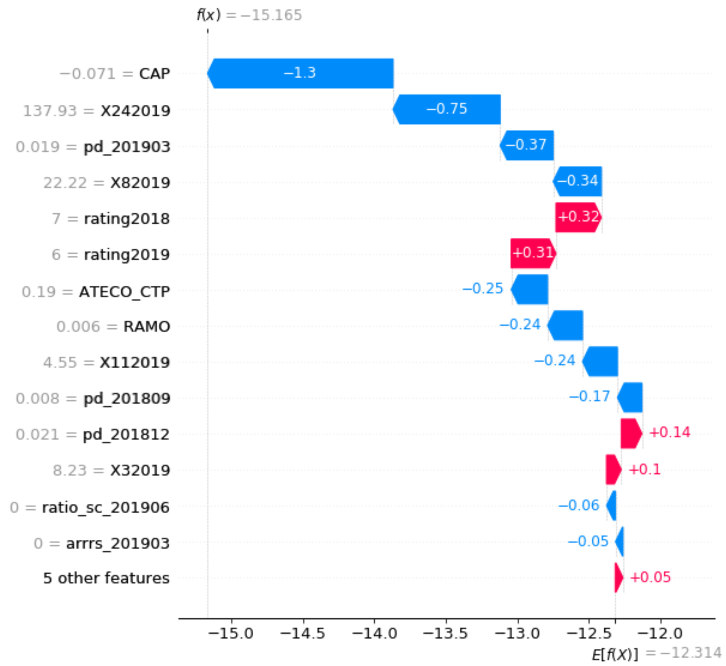


Figure 5.5: The highest SHAP values of the features for a single negative (not bankrupt) prediction by the **CAT** classifier. In this particular example is relevant the detailed geographical localization plus a couple of balance sheet indicators ($X24$ and$X8$) and a good value for the probability of default. To be noted, instead, the negative impact of a bad rating classification.

5.4 Adjusted default prediction

Regarding Adjusted default prediction, we first consider the prediction exercise we conducted using both the credit datasets from Central Credit Register (CCR) and AnaCredit (i.e. Mixed dataset). An analysis of the results obtained using the **CAT** classifier are shown in the following figures. Also in this case if we use other classifiers we obtain similar results.

We can see (Fig. 5.6) that the most important factor in the explanation of prediction are the incidence of overdraft (sc) and margins (m) for the last reference dates available. These two variables belong to CCR dataset, but it is interesting to note that also some features from AnaCredit dataset show an impact on overall explainability, for example arrears (arrrs) and probability of default (pd). This analysis corroborates our conjecture that in order to predict adjusted default, the combination of the two credit datasets (CCR and AnaCredit) can improve the results. We found a similar evidence regarding the performance prediction and now we have a confirmation with reference to the ability to explain the results.

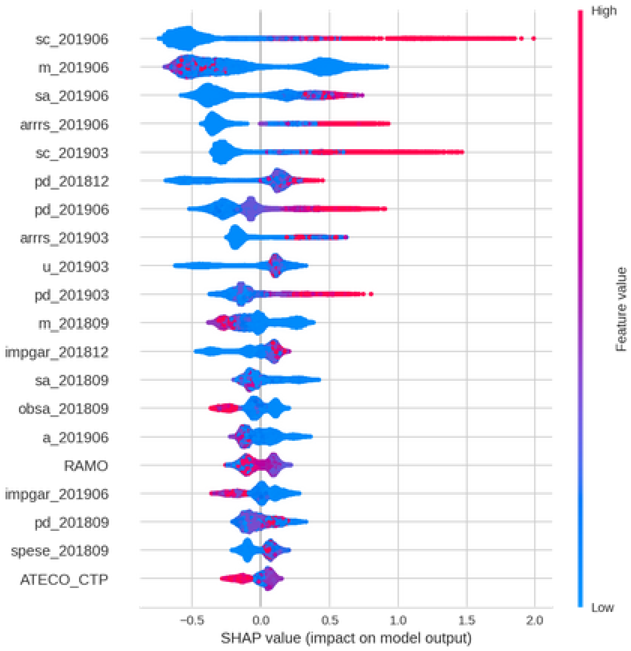


Figure 5.6: Explanatory overview with **CAT** classifier. The figure shows a general overview of the main driver factors of the predictions about the use of MIXED dataset. The most important factors in the explanation are overdraft (sc) and margins (m) from CCR dataset, and arrears (arrrs) and probability of default (pd) from ANACREDIT dataset.

Finally, in Fig. 5.7 we consider the prediction with MERGED dataset. We observe that also for adjusted default prediction the detailed geographical classification (CAP) represents the predominant factor. Subsequently, some important characteristics belonging to all three elementary datasets appear relevant. We refer to overdraft (sc) from CCR dataset, revenues (X1) from BALANCE dataset, arrears (arrrs) and probability of default (pd) from ANACREDIT dataset. As in the previous section in Fig. 5.7 we are considering a overall picture of the factors that can explain the prediction. But the relevant added value resulting from the use of SHAP lies on the capability to provide explanation for each single prediction.

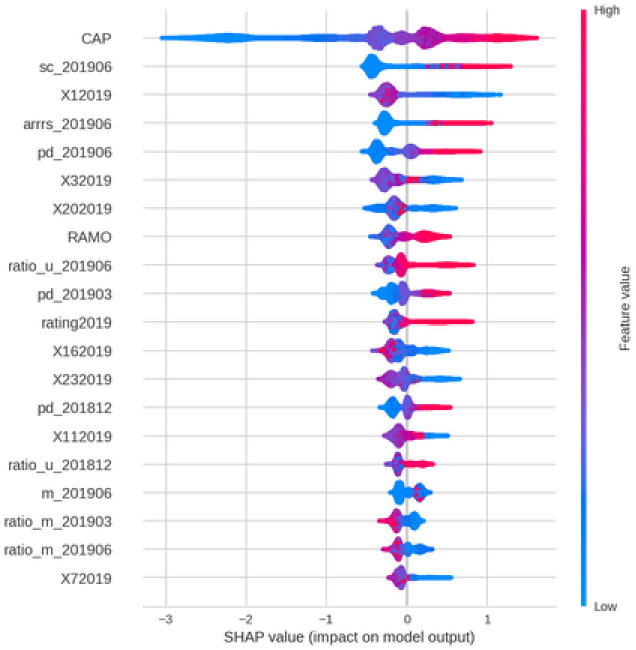


Figure 5.7: Explanatory overview **CAT** classifier. The figure shows a general overview of the main driver factors of the predictions with Merged dataset. Detailed geographical classification (CAP) represents the predominant factor. Subsequently, some important characteristics belonging appear relevant: overdraft (sc) from CCR dataset, revenues (X1) from Balance dataset, arrears (arrrs) and probability of default (pd) from AnaCredit dataset.

5.4.1 Explainability of single prediction

As for the bankruptcy prediction, we show how the Shap values for both healthy and failed companies provide a clear picture of the reasons why that prediction has been made, accordingly to the explanatory overview. In Fig. 5.8 we can see an example of positive prediction (i.e. a prediction of a future default). In this particular case, the driver factor of a default prediction is connected with the detailed geographical classification (CAP) but also with a very negative rating classification and with a bad value probability of default (pd). In the last example (Fig. 5.9) we consider a negative (no default) prediction. In this particular case a good classification regarding probability of default and rating are the principal factor that explain the no default prediction, while geographical classification (CAP) slightly pushes the prediction in the opposite direction. We can observe that also the liquidity condition (X16) of the firm play a significant role in prediction.

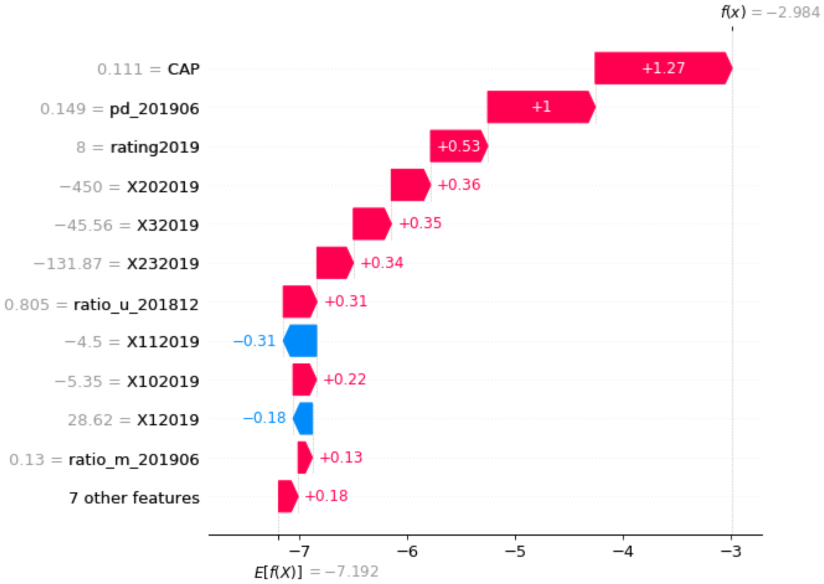


Figure 5.8: Explanation of a single positive (default) prediction using Merged dataset. Also in this case we can observe a significant impact of the detailed geographical classification (CAP). Others relevant factors are probability of default (pd) and rating for the last available reference date.

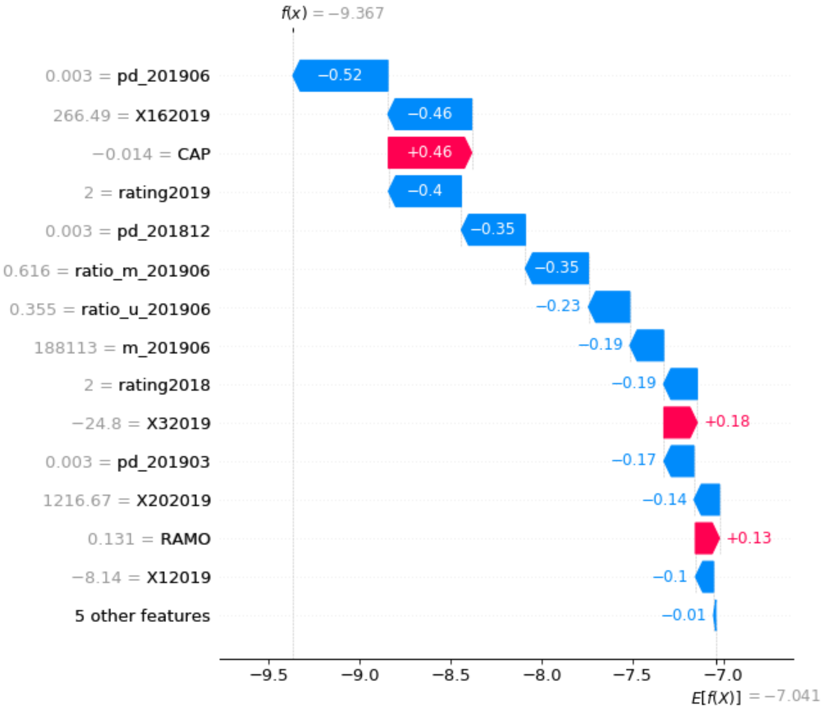


Figure 5.9: Explanation of a single negative prediction using Merged dataset. In this example a good classification regarding probability of default and rating are the principal factors that explain a negative (no default) prediction, while geographical classification (CAP) slightly pushes the prediction in the opposite direction. Also the liquidity condition (X16) of the firm play a significant role.

5.5 Lessons from Explainability Analysis

In this section we want to highlight the main results of our analysis relating to the explainability both with reference to the prediction of bankruptcy and bank default.

1. A first result regards the complementarity of the datasets that we used in a combined manner. We refer to the MERGED dataset that combine all the three elementary datasets. In fact, both for bankruptcy prediction and for adjusted default prediction, we observe that the most important features belong to all the three datasets involved (see, for example, Fig. 5.2 and Fig. 5.7). This corroborates our conjecture that in both cases the combination of the elementary datasets improves performance. The explainability of the prediction results confirms the complementarity by showing us that there are relevant features in all the three datasets. Our conclusion is that, in general, the combination between balance-sheet data and credit data provide a significant gain in the ability to prediction of firms default.

2. A second result concerns with the clear identification of some predominant features. In particular we refer to: rating, probability of default, overdraft, arrears. These features come from both the balance sheet dataset and the two credit datasets and are very important in order to predict both bankruptcy and bank default. With particular referenc to balance sheet data, the features that are most relevant in the prediction suggest a clear role of well known indicators (ROE) together with some information often connected with the revenues and liquidity of the companies, which would therefore seem to constitute a possible usable signal of crisis.

3. Third result: the manipulation of some features is extremely useful to enhance the explainability, even if it does not provide significant gains in the performance of predictions. In particular, we refer to the pre-processing of categorical features and to the synthetic features we created (see Section 3.1). Regarding the latter, ratio appears much more relevant respect the changes over time. This point could be linked with the Section 4.3 in which we provide an analysis of the impact of historical depth of data on predictive performance.

4. Last point: what we expected and what surprised us. We expected some important indicators calculated specifically to assess the condition of companies (for example, rating and probability of default) to be useful for the prediction. We also expected economic signs of corporate weakness: overdrafts (sc), arrears in payment (arrrs) to play a role. This has been confirmed by our experiments but perhaps with even greater prominence than we thought. Instead, we expected that features built taking into account the evolution over time could play an important role but it would not seem so. In fact, as we highlight in Section 4.3, the information that appears most relevant is those relating to the most recent reference dates, while past information seems to be much less relevant. This result was also not entirely expected.

Chapter 6

Prediction performance analysis in an unbalanced scenario

In this chapter we study the context of prediction in an unbalanced scenario with a more detail. Our objective will be an improvement in default prediction, also with a particular regards to the stakeholder needs.

6.1 Heavy imbalanced dataset

A lot of modern problems in the machine learning field have to deal with imbalanced datasets. For some of them, the problem is due to the lack of data samples, for others to the intrinsic nature of the problem.

In our case, the problem falls into the second category since we are trying to assess the likelihood of companies to fail and just a small amount of the total firms fail over the year. For our datasets the ratio between failed and healthy companies is very low (less then 5 % in 2015 with a decreasing trend in the following years).

In the following, we investigate the possibility of significantly varying some performance indicators of the prediction results in function of the imbalance of the dataset. In fact, the use of a balanced training set in prediction lead to the maximization of Recall (TPR) and True Negative Rate (TNR), with the consequent minimization of Type-I and Type-II errors. On the other hand, the use of a highly unbalanced training set can significantly increase Precision.

We have already addressed this issue in our experiments in which, in particular, we carried out the classifications using both a fully balanced training set and an unbalanced training set that maintains the same imbalance of the test set. We have already shown on that occasion the differences found for the different performance indicators we used.

In this chapter we try to explore better this point in order to improve prediction performance in highly unbalanced contexts.

First of all, we show in Fig. 6.1 the dynamic of some performance indicators when we change the imbalance of the training set. In this experiment our dataset has an imbalance of 4% (percentage of defaulted firms over the healthy firms). We perform our prediction with one of the best classifier (LGBM) and using a variable training set. In particular, we consider on the x-axis the percentage that indicates the ratio of defaulted firms in the training set respect to the healthy firms. The percentage of 100% indicates a perfectly balanced training set while that of 4% represents the original imbalance of our dataset (which corresponds with the imbalance of the test set). But to get a complete picture of the trends of the performance indicators, we also tried to use percentages of failed companies that go beyond 100% and below 4%. In Figure 6.1 we can see a steep increase in Precision as the training set imbalance increases. On the other hand the gain in Precision comes at the expense of a sharp reduction in Recall. It is interesting to note the different trends for the two main relevant performance indicators we are considering. AuROC grows if we increase the percentage of positive companies in the training set and reaches its maximum value for a perfectly balanced training set. Furthermore, the performances remain excellent even if the failed companies are greater than the healthy ones up to a percentage of 200%. Instead, F1-score reaches a maximum at the point where Precision and Recall assume the same value, while for a perfectly balanced training set the F1-score value is lower.

In the previous section 3.1.2 we provided a brief overview about the principal techniques typically used in case of unbalanced dataset while in the following (Section 6.2 and the subsequent ones) we will deepen our analysis on predictions in strongly unbalanced contexts.

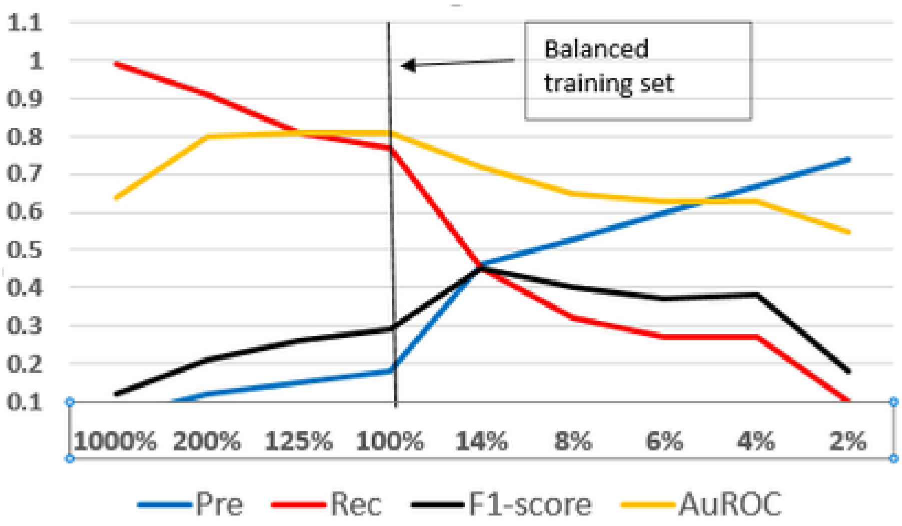


Figure 6.1: Analysis of prediction performance indicators as the training set imbalance changes. The percentage on the x-axis indicates the ratio of defaulted firms respect to the healthy firms in the training set. We can see a steep increase in Precision as the training set imbalance increases. On the other hand the gain in Precision comes at the expense of a sharp reduction in Recall.

6.2 Can we try to do something more? An insight into measuring results

In the experiments reported in Chapter 4 our main objective was to find the best possible performance in prediction. In particular, we obtained our best performance in prediction

using a perfectly balanced training set and measuring performance using the AuROC. As we discussed before, this represents a choice widely shared by the most important literature on the subject, in which generally we can observe a tendency to maximize the Recall (TPR) and try to minimize Type-1 and Type-2 errors (see for example [7]). This result is typically achieved using a perfectly balanced training set, which guarantees a high Recall value but at the expense of a lower Precision.
As we can observe in Fig 6.1, the performance of predictions varies considerably when an unbalanced training set is used. For example, in our first set of experiments reported in Chapter 4 we used both a training set that maintains the same imbalance as the original dataset and a perfectly balanced training set (see also our first paper on the subject [3]).

Moreover, also the choice of the performance indicator can be relevant in the evaluation of the results. In [45] an accurate analysis about the performance of predictions is carried out in an unbalanced context, considering also a dynamic variation of the training set. In that paper, a proposal of cost function is also introduced. It takes into account the gain deriving from correct predictions of the minority class and the losses arising from the errors in classification. The conclusion in that paper show that a balanced training set is not always the optimal choice, but it is only if the gain for a correct prediction (on the minority class) is much greater than the loss for an incorrect prediction (always on the minority class). As we anticipate in Chapter 4, we try to use this paper as a reference point in our following analysis. In particular, in section 6.3 we will extend the approach in [45], however, suggesting a new proposal of linear gain function that takes into account both the results obtained on the minority class and those relating to the majority class. In other word, our gain function will take in account all the four component of the confusion matrix: TP (Tue Positive), TN (True Negative), FP (False Positive) and FN (False Negative).

Our goal will be to identify a structured framework to apply to bank default prediction that helps to maximize the gain obtained, according to some predetermined criteria identified on the business side.

More in detail, we will follow the steps below:

- First of all, we will extend the analysis of the variation in predictive performance as a function of the imbalance of the training set that we introduced earlier (see Fig. 6.1).
- We will propose a linear cost function that takes into account all the four components of the confusion matrix and consider the stakeholder point of view, in connection with default predictions.
- Finally, we will draw some conclusions oriented to strengthen the effectiveness of the default prediction activity in an highly unbalanced context.

6.2.1 Prediction in a highly unbalanced context: which performance measure is best to use?

As we said before, in order to measure the prediction results in many important studies it is used AuROC in combination with a perfectly balanced training set. It is also the main choice in our work in order to evaluate the performance of the ML classifiers and to assess the explainability of the predictions. One of the reasons behind this approach release in connection with a wider possibility to evaluate prediction performances using a comparison with similar works.

But in this chapter we try to discuss to what extent the use of AuROC represents always a good choice and even if it is the best choice to use a perfectly balanced training set in default prediction.

First of all, we can reason whether a high AuROC value always represents an excellent performance in a strongly unbalanced scenario. In fact, we can start the reasoning observing that very different classification performances can match to the same value of AuROC. In order to analyze this point, we can consider a binary classification over a sample of 1000 total elements (960 negative and 40 positive elements).

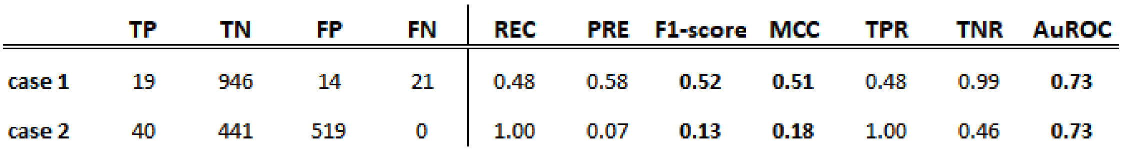
In Table 6.2 we show two very different cases though they have the same AuROC value. But we observe that they represents two really different classification results (19 True Positive for case 1 versus 40 True Positive for case 2, while 946 True Negative for case 1 versus 441 for case 2: what is better?).

In addition, we can mention that the two case obtain very different F1-score values (and also very different MCC values).

What is the best performance? The reply to this question is not always simple. It depends in a significant manner on whether we are interested in predicting the minority class only or whether we are also interested in predicting the majority class.

For example, if we are mainly interested in the correct identification of the minority class, the case 2 could be the best option. In this case we obtain the correct identification of all the minority class: 40 True Positive classifications over 40 of total positive elements. On the other hand, in this case we have identified less than half of the majority class: only 441 compared to a total of 960 negative elements.

Observing the performance indicators in Fig. 6.2, we have a further confirmation that F1-score more rewards a balanced stance between Precision and Recall, while AuROC rewards more the correct prediction of the minority class (and therefore a high Recall). This conjecture is perfectly in line with our previous findings reported in Fig 6.1.



	TP	TN	FP	FN	REC	PRE	F1-score	MCC	TPR	TNR	AuROC
case 1	19	946	14	21	0.48	0.58	**0.52**	**0.51**	0.48	0.99	**0.73**
case 2	40	441	519	0	1.00	0.07	**0.13**	**0.18**	1.00	0.46	**0.73**

Figure 6.2: Confusion matrix and performance indicators for two different binary classifications. We have considered a sample of 1000 total elements highly unbalanced (960 negative and 40 positive). The two different classification cases (case 1 and case 2) have the same AuROC value but they represents two really different classification results (19 True Positive for case 1, versus 40 for case 2. On the other hand 946 True Negative for case 1 versus 441 for case 2). Moreover, we can observe that the two cases show very different values for F1-score and MCC performance indicators.

In the following we continue to consider our simple dataset of 1000 elements strongly unbalanced at 4%, with 960 negative and 40 positive elements. In particular, here we want to analyze the differences and the relationships between some widely used performance indicators.

In Fig. 6.3 we show the relation between F1-score and AuROC for all the possible classification combinations in the case of 1000 total elements, with a minority class of 40. In other words, we considered all the possible values for the confusion matrix (i.e. 960 ∗ 40 different combinations) and calculated the correspondent performance indicators. The figure shows the scatter plot related to F1-score and AuROC values for all the possible classification points. We can observe that exists a large interval of F1-values in correspondence with a single value of AuROC. This means that they exist a lot of cases for which we obtain the same AuROC value but we obtain at the same time significantly different F1-score values (like in the case 1 and case 2, in Fig. 6.2). Moreover, we observe that this fact is particularly true for high value of AuROC (from 0.6 upwards). In addition we found that, to a lesser extent, the opposite situation is also true: in fact, we observe classification points that have the same F1-score values but different values for the AuROC.

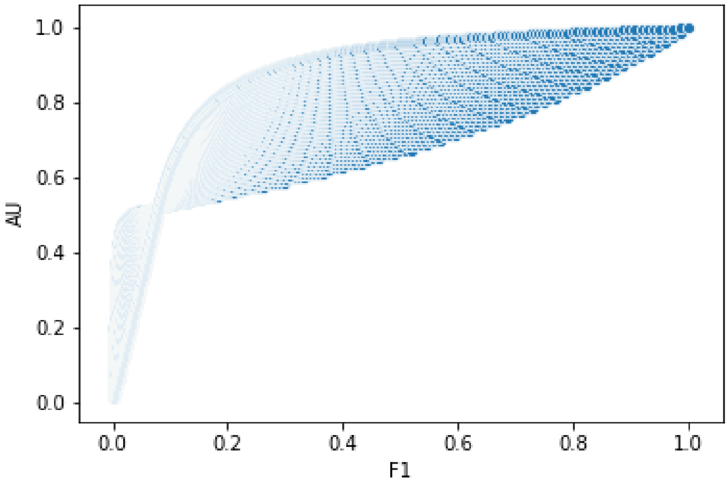


Figure 6.3: Scatter plot that reports the values of AuROC (y-axis) and F1-score (x-axis) for all the possible classifications for the sample considered. We have considered a simple dataset of 1000 total elements highly unbalanced (960 negative and 40 positive elements).

In the following figures we report other scatter plots between some pairs of performance indicators. In these experiments we consider two different scenarios for each couple of indicators: a balanced dataset (on the left of each of the following figures) and an unbalanced scenario (on the right), similar to the example we mentioned before.

We can observe that F1-score and MCC show a similar behaviour when we consider a unbalanced dataset (Fig. 6.4, on the right) while AuROC and MCC have a good correspondence in balanced scenario (Fig. 6.5, on the left). Instead, in the other cases there are many classification points in which we observe the same phenomenon described above. More precisely, we observe many different cases in which for one indicator we achieve the same performance result while they show very different performances results for the other indicator we are comparing.

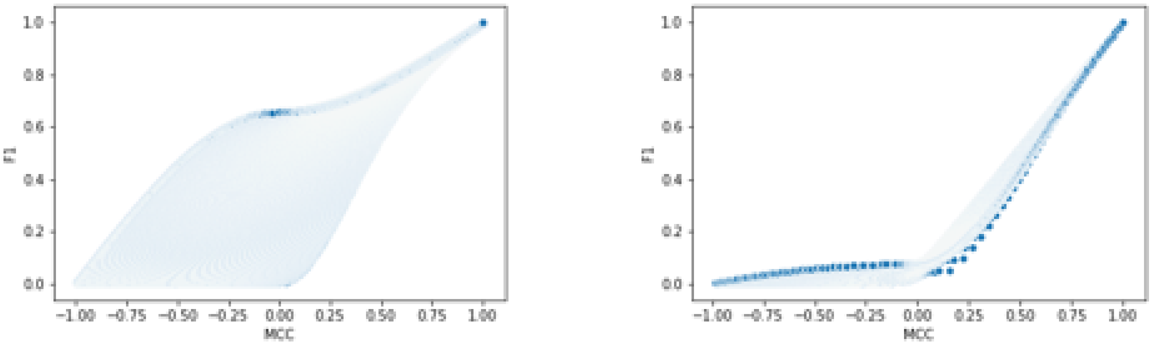


Figure 6.4: Scatter plot between F1-score and MCC. In the figure we can observe two scatter plot graphics that reports the values of F1-score (y-axis) and MCC (x-axis) for all the possible classifications for the dataset considered. On the left we can observe the results for a balanced dataset of 1000 elements with 489 positive, while on the right we consider a highly unbalanced dataset of 1000 elements with 40 positive elements.

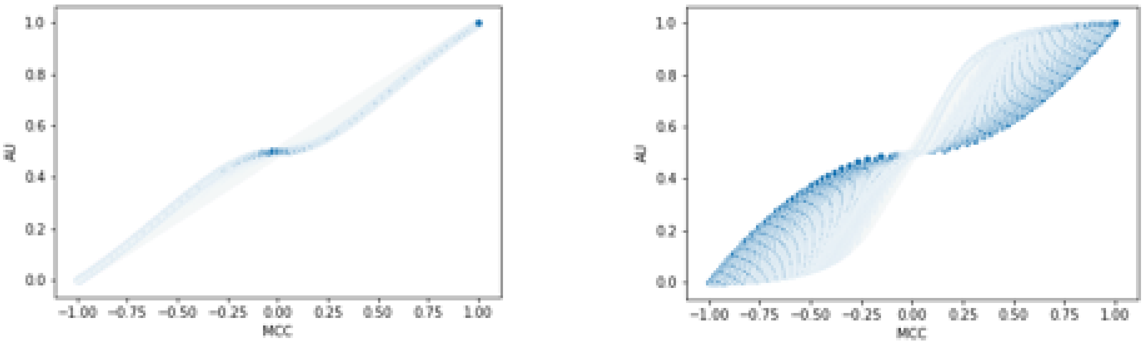


Figure 6.5: Scatter plot between AuROC and MCC. In the figure we can observe two scatter plot graphics that reports the values of AuROC (y-axis) and MCC (x-axis) for all the possible classifications for the dataset considered. On the left we can observe the results for a dataset of 1000 elements with 489 positive, while on the right we consider a dataset of 1000 elements with 40 positive.

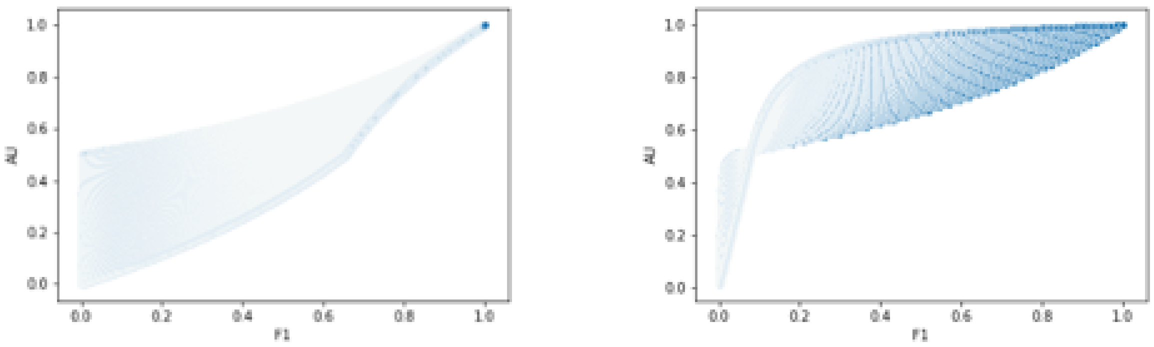


Figure 6.6: Scatter plot between AuROC and F1-score. In the figure we can observe two scatter plot graphics that reports the values of AuROC (y-axis) and F1-score (x-axis) for all the possible classifications for the dataset considered. On the left we can observe the results for a dataset of 1000 elements with 489 positive, while on the right we consider a dataset of 1000 elements with 40 positive.

To summarize, we can assert that the choice of the performance indicator is certainly not irrelevant in measuring the results.

6.2.2 Which training set should we use?

In this section we are interested to better understand what kind of training set should we use in order to maximize the prediction performance. In particular, we would like to evaluate whether a completely balanced training set can be consider the best solution. In the following we extend the analysis we started in Section 6.1 performing some experiments that involve two datasets with a different degree of imbalance. We consider the prediction results as the training set varies, starting from a completely balanced scenario (training set with $ratio = 1$ between defaulted firms and healthy firms). Then, we gradually move to an increasingly unbalanced training set up to the natural imbalance of the overall dataset, that is equivalent to the imbalance of the test set.

For these experiments we use LGBM algorithm, without making special settings on the parameters of the classifier.

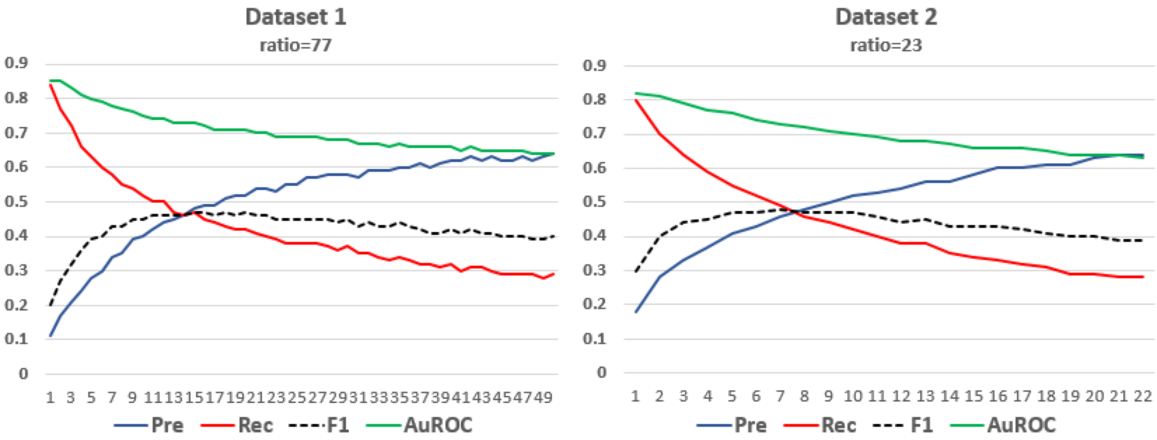


Figure 6.7: Analysis of the variations of some performance indicators as a function of the imbalance of the training set used for the predictions. The numbers reported on the x-axis represent the imbalance of the training set. We consider two different datasets: Dataset 1 (about 570,000 rows, on the left) and Dataset 2 (about 300,000 rows, on the right) that have two different degree of imbalance: the ratio between the majority class (no default) and the minority one (default) is equal to 77 and 23, respectively. In these experiments we perform the predictions using LGBM classifier.

We can observe the results in Figure 6.7: AuROC is maximum for a completely balanced training set and decrease as the imbalance of the training set increases. We already seen similar results in Fig. 6.1 and so, as we already observed, Precision and Recall show opposite trends. More precisely, Recall is maximum when the training set is fully balanced and decreases significantly as the train set imbalance increases. Instead, F1-score shows a maximum point corresponding to the scenario in which Precision and Recall assume the same value.

Our results regarding the dynamic of Precision, Recall and F1-score are absolutely in line with those reported in [45]. However, we perform the prediction using two different datasets with a different degree of imbalance. So, we can add the observation that the degree of imbalance of the training set for which F1-score reaches its maximum seems to depend on the overall imbalance of the dataset. That is, in other words, equals to the unbalance of the test set we used. In fact, we can see (on the left in Fig. 6.7) that the point at which Precision and Recall assume the same value falls in correspondence with an imbalance equal to about 15 (ratio between majority class and minority class in the training set). In this case the imbalance of the test set is equal to 77. On the other hand, when the imbalance of the test set is equal to 23 (on the right of Fig. 6.7) the point at which Precision and Recall assume the same value falls in correspondence with an imbalance for the training set equal to 7.

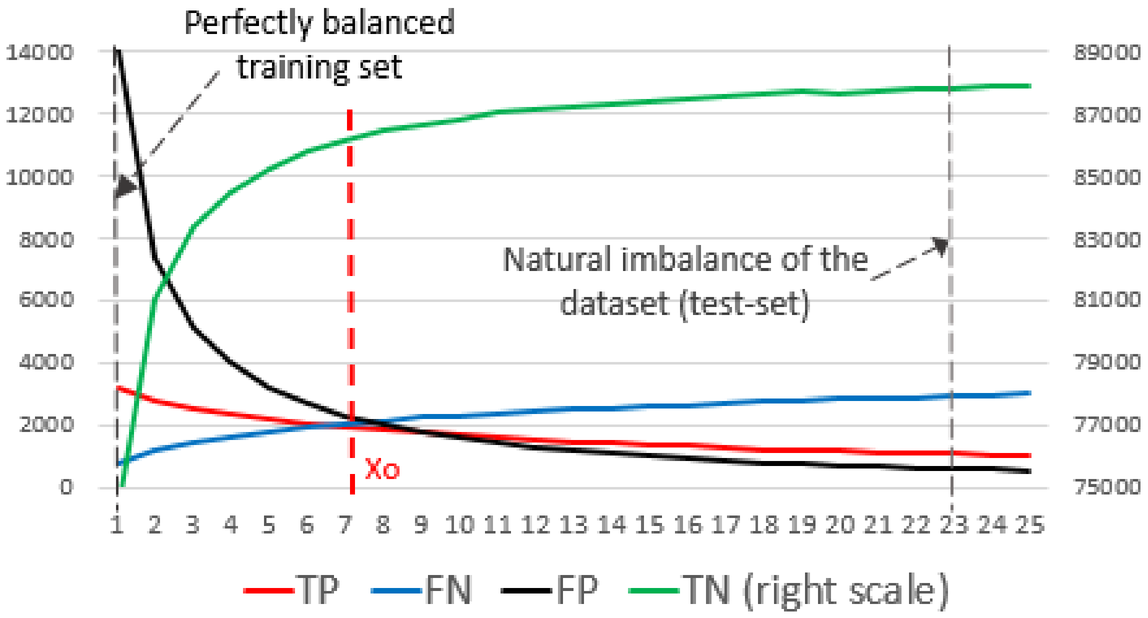


Figure 6.8: Confusion matrix changes: Variation of number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) in function of the imbalance of the training set. The red dotted line highlights a change in slope for the TN curve (and consequently for FP curve).The numbers reported on the x-axis represent the imbalance of the training set (i.e. ratio between majority class and minority class).

In Figure 6.8 we show the variation of the confusion matrix in function of the training set imbalance, for the second dataset we used before. In particular we recall that the dataset has an imbalance equal to 23. But we clarify that the results we obtained are the same also for the more unbalanced dataset, in terms of dynamic of the several components of the confusion matrix.

We can observe that the number of True positive (TP) is maximum using a balanced training set but it decrease sharply if we increase the training set imbalance. On the contrary, the number of True Negative (TN) shows an exactly opposite dynamic, even in terms of absolute number they are very different (since the negative class represents the majority one). The slopes of growth [decrease] of FN [of TP] are reduced only slightly instead as the imbalance of the training set increases. It is interesting to note that we found a value of the training set imbalance for which the TN [FP] growth [decrease] slope is drastically reduced. Therefore, to the left of this point on the graph, increasing the imbalance results in much more TN and much less FP. To the right of this point an increase in imbalance will impact these variations much less. This could be relevant in the search for an optimal prediction performance connected with how relevant we consider to obtain more TN at the expense of less TP.

To conclude, according to the observed trend of the four components of the confusion matrix as the unbalance of the training set varies, we can assert that the overall correct predictions (TP+TN) are lower using a completely balanced training set while the overall prediction errors (FP+FN) are larger. This result (highlighted in Fig. 6.9) arise as a consequence of the different slope of the TP and TN curves (and correspondingly FP and FN curves) that we observed in Fig. 6.8 as we reduce the imbalance of the training set and tend towards a perfectly balanced training set approach (left part in Fig. 6.8).

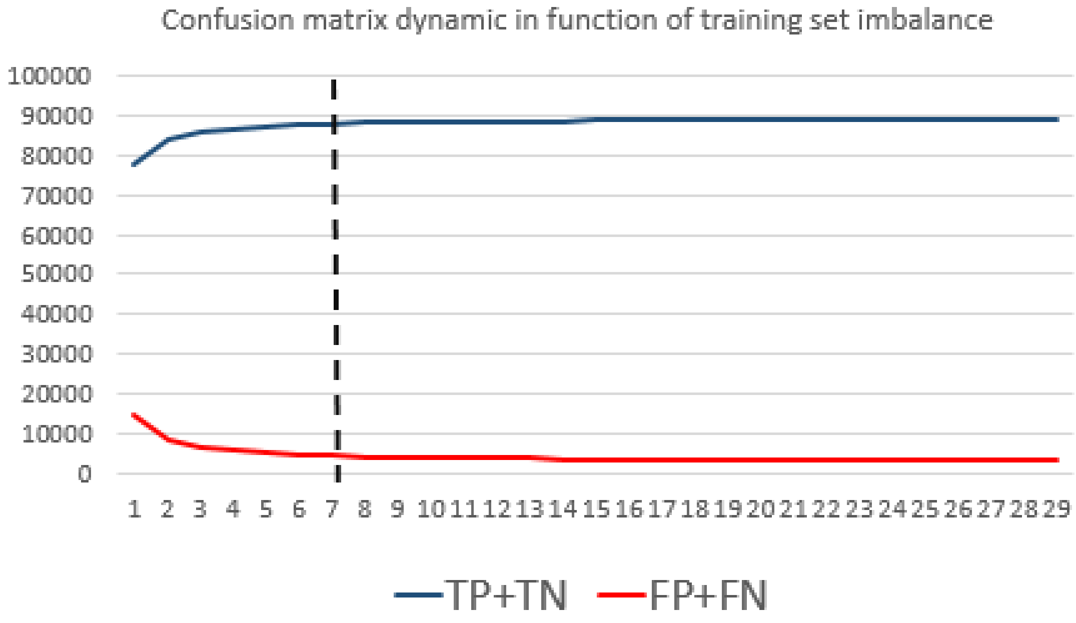


Figure 6.9: Analysis of the variations of the confusion matrix as a function of the imbalance of the training set used. The imbalance (x-axis) is expressed as the ratio between the majority class (no-default) and the minority class (default). In the left part of the graph we can observe that when we use a completely balanced dataset we obtain an overall reduction of the correct predictions (TP+TN curve) and a simultaneous increase of the errors in prediction (FP+FN curve).

6.2.3 Some consideration about performance indicators

In this section we would like to draw some consideration about the use of the performance indicators. As we known, the evaluation of the prediction performances presented in the previous chapters was mainly based on some well-known performance indicators. In particular we start using a set of important indicators such as AuROC, F1-score, TPR, TNR and others; then, in the second section of results chapter we mainly focused our attention on AuROC in line with with a large part of the most relevant literature. In the previous sections of this chapter we try to disentangle the results of the different performance indicators in connection with the imbalance of the dataset and, in particular, the imbalance of the training set that we use. Then, we demonstrate that the choice of the performance indicator is not irrelevant in the evaluation of the performances.
To summarize, we found that:

1. Different performance indicators show different results and it is difficult in some cases to decide the best results. In Figure 6.2 we can observe two different examples of binary

classification that have very different performance in terms of F1-score and MCC, while they have the same AuROC value.

2. We extend the analysis related to the dependence of some performance indicators from the confusion matrix. In particular, we observe in Fig 6.8 that if we reduce the imbalance of the training set and we tend towards a fully balanced scenario AuROC increases while F1-score decreases.

3. In Fig. 6.9 we show that if we reduce the imbalance of the training set and consequently we tend towards a fully balanced scenario, the total correct classifications (TP+TN) decrease while the total errors (FP+FN) increase.

The issues summarized above invite us to a further reflection regarding performance measurement and the use of a fully balanced training set. In particular the third point would suggest that the performance of the prediction is worsening when we tend toward a completely balanced training set, although we have seen that the value of the AuROC increases.

To simplify, we could conjecture that using a balanced training set makes us get a better AuROC but at the same time a worse F1-score. But, if we just simply count correct classifications and errors, the situation seems to worsens.

The last point that we want underline in this section regards the fact that all the performance indicators we used take into account only information contained in the confusion matrices that result from the classification procedures.

In the next section we try to consider also an other point of view, given that our prediction are related to an important business problem. The attempt will be to incorporate in the evaluation of the results also the business point of view.

6.3 A model to maximize the value of predictions

In this section we propose a linear gain function (Total Gain, GT) that take into account all the four components of the confusion matrix. We define the total gain function GT according to the following formula:

$$GT = \alpha * TP + \beta * TN - \gamma * FP - \delta * FN \tag{6.1}$$

where $\alpha, \beta, \gamma, \delta$ are positive values that we can use in order to determine the relevance of each components of the confusion matrix.

This proposal extends the gain function introduced in [45], in which are used only TP and FP. Therefore, in that paper the author suggest exclusive relevance only to the prediction results relating to the minority class. This approach is reasonable but it implies the risk of excessively rewarding the achievement of correct positive classification (TP), if these are considered more important than errors on classification of the minority class (FP). But we observe (Fig. 6.8) that when we tend towards a balanced training set we obtain an increase of TP but at the same time a greater reduction (in absolute number) of TN. Hence our choice to consider all the four different components of the confusion matrix in our gain function GT.

Taking into account that:

- $P = TP + FN$ (total positive)
- $N = TN + FP$ (total negative)

We can rewrite equation 6.1:

$$GT = (\alpha + \delta) * TP - (\beta + \gamma) * FP + \beta * N - \delta * P \tag{6.2}$$

It is interesting to analyze the dynamic of GT function when we modify the unbalance of the training set. In particular we refer to Fig 6.8 and consider a variation in the imbalance of the training set, that means a different point on the x-axis in the graph. The correspondent variation ΔGT of the gain function is:

$$\Delta GT = (\alpha + \delta) * \Delta TP - (\beta + \gamma) * \Delta FP \tag{6.3}$$

This means that if we go to left in the graph, towards a balanced training set, the total gain GT will increase only if:

$$(\alpha + \delta) * \Delta TP > (\beta + \gamma) * \Delta FP \tag{6.4}$$

But, as we saw in Fig. 6.8, the ratio ($\Delta FP/\Delta TP$) is very high in the left part on the graph, increasing as we tends to a completely balanced training set scenario.

So, we can write that:

$$\Delta GT > 0 \leftrightarrow \frac{(\beta + \gamma) * \Delta FP}{(\alpha + \delta) * \Delta TP} < 1 \tag{6.5}$$

Since $\frac{\Delta FP}{\Delta TP} >> 1$, our conclusion is that it is convenient to use a fully balanced training set if and only if the following condition holds:

$$(\alpha + \delta) \gg (\beta + \gamma) \tag{6.6}$$

But, this appear a really residual scenario, in all the other case we conclude that it is useful to increase the imbalance of the training set in order to improve the performance of the prediction.

Moreover, we can observe immediately that the previous scenario coincides with the case:

$$\alpha \gg \gamma \tag{6.7}$$

if we assume that

$$\delta \simeq \beta \tag{6.8}$$

This conclusion is perfectly in line with the results in [45] in which, in fact, they are not considered the relevance of prediction about the majority class.
But, as we anticipate before, we think that it is not a effective approach to consider irrelevant the contribution of the correct True negative (TN) classification and the negative impact of the False negative (FN) classification (that is the same as saying $\delta = \beta \simeq 0$).
In addition, we consider that the condition reported in 6.7 and, more in general, also the condition in 6.6 could be represent in many case a too strong requirement. For example, in the case of a bank that would increase its credit portfolio or in a case in which a Public Authorities would support the economy by issuing guarantees to firms in order to obtain new loans.

6.3.1 Some case studies

In this section we try to fix some more general consideration about the measurement of prediction performance. Our starting point could be the observation that the use of different performance indicators does not always clarify which are the best performances. The matter becomes particularly complicated when we operate in a highly unbalanced context.

One relevant question could be: "Are we only interested in performance on the correct classification of the minority class?"

In this chapter (Sect. 6.3) we try to deal with issues introducing a gain function GT that take into account both minority and majority class (and both in terms of correct classification and errors).

An extreme case could be that in which we give importance only to correct positive classification and we want avoid false negative classification (i.e to miss a positive case). This means that we consider preponderant in GT function (6.1) the values of α and δ. As we can already state in condition 6.6 this implies that balanced training set can maximize the overall gain.

In the following, we report four case studies, that represent four different examples relating to our reference context concerning the business failures. The choices we made in setting the parameters of the gain functions GT (and the relative business justifications) are absolutely questionable. But the main objective of our exercise is to show, however, how the business-side evaluation related to the convenience of the predictions can significantly change the final gain obtained from them. In fact, we think that the more effective way to measure a prediction performance framework should take into account the business assessments of the stakeholders involved in the predictions and should reflect their needs.

In addition, our practical cases will confirm the conjecture that to use a perfectly balanced training set is not necessarily always the best choice, even it leads to predict a much higher number of positive cases.

Example 1: Point of view of a private bank that gives a new loan

- $\alpha = 0$: to identify correctly a future default implies that the bank do not extend a new loan

- $\beta = 10$: to identify correctly a no-default determine a gain (we assume 10%) for the bank that extend a new loan of 100
- $\gamma = 10$: to have a FP determine a miss gain (10%) for the bank that doesn't give a loan of 100
- $\delta = 100$: to have a FN implies a loss for the bank that give a loan to 100 to a future default firm

We can observe in Fig. 6.10 that if we use a perfect balanced training set we obtain a significantly lower total gain.

Example 2: Point of view of a private bank that manages its loan portfolio

- $\alpha = 50$: to identify correctly a future default implies that the bank can act some measure in order to reduce losses, let's assume 25% of the total portfolio
- $\beta = 10$: to identify correctly a no-default determine a gain for the bank that can extend new loan to its client
- $\gamma = 20$: to have a FP determine a miss gain for the bank that act measure to contain the risk and doesn't give a new loan to its client
- $\delta = 100$: to have a FN implies a loss for the bank that can loss all the asset due to the default of the firm

We can observe also in this case (Fig. 6.10) that it isn't convenient to use a perfectly balanced training set.

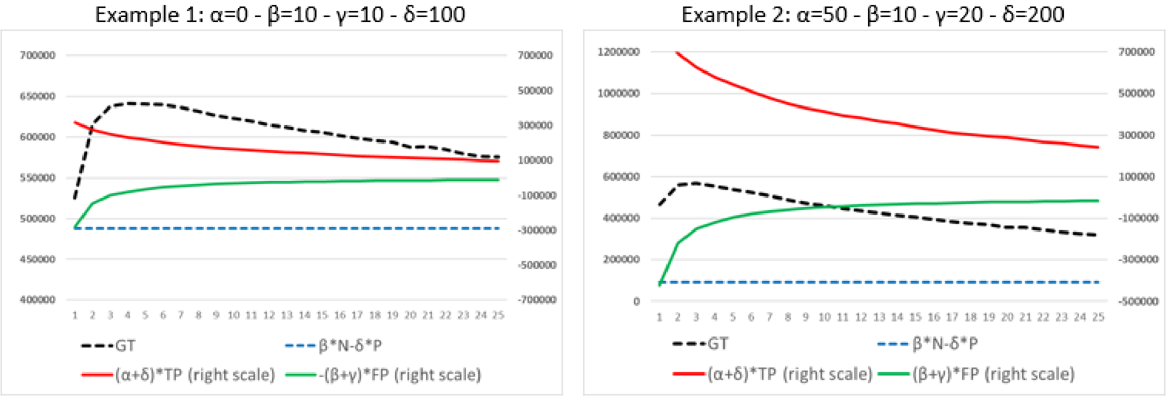


Figure 6.10: Example 1 (on the left): Point of view of a bank that gives a new loan. Example 2 (on the right): Point of view of a bank that manages its loan portfolio. In this two examples we consider the dynamic of the gain function GT and its component in equation 6.1.

Example 3: Point of view of bank Supervisory Authority

- $\alpha = 100$: to identify correctly a future default implies that the supervisor can act measure to mitigate bank's risk obtaining a gain in ensuring overall financial stability

- $\beta = 0$: to identify correctly a no-default doesn't implies a gain for the Supervisor perspective
- $\gamma = 0$: to have a FP determine a not necessary measure imposed on the bank (for example additional capital)
- $\delta = 150$: to have a FN implies a case for which supervisor does not take into account the individual risk of the bank by exposing the entire system to serious risks

We can observe in Fig. 6.11 that in this case a fully balanced training set assure the maximum total gain GT.

Example 4: Point of view of bank Supervisory Authority (with a little integration)

- $\alpha = 100$: to identify correctly a future default implies that the supervisor can act measure to mitigate bank's risk obtaining a gain in ensuring overall financial stability
- $\beta = 0$: to identify correctly a no-default does not implies a gain for the supervisor
- $\gamma = 30$: to have a FP determine a not necessary measure imposed on the bank (for example additional capital). But we can assume that this measure represent a loss for the bank profitability and can also involve risks for the entire financial system
- $\delta = 150$: to have a FN implies a case for which supervisor it does not take into account the individual risk of the bank by exposing the entire system to serious risks

We can observe that the little integration in the treatment of FP coefficient (compared to the previous case) determines it's no longer convenient to use a perfectly balanced training set (Fig. 6.11).

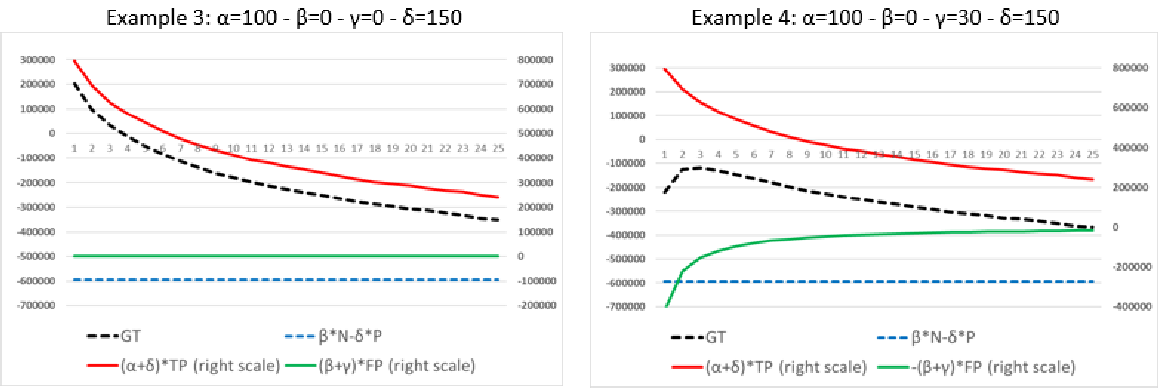


Figure 6.11: Example 3 (on the left): Point of view of bank supervisory Authority. Example 4 (on the right): Again the point of view of bank Supervisory Authority but with a little integration. In this two examples we consider the dynamic of the gain function GT and its component in equation 6.1.

As we can see in the previous case studies, only one time the gain function (GT, dashed black line in the figures) reach its maximum value in correspondence of a fully balanced training set scenario (Example 3, on the left in Fig. 6.11).

In fact, we can note that in this case we met the conditions indicated in the formula 6.6. In other words, in this scenario we are giving mainly relevance to correct positive classification and to avoid false negative classification. This means that we consider more important in GT function (6.1) the values of α and δ. But this approach pay also a cost in terms of less number of True negative (TN) and a major number of False positive (FP). In particular, for each single correct classification in the minority class (TP) that we gain, we pay a high cost that means to obtain a higher number of errors (FP). But, if we let's take it to the extremes, we may be tempted to classify every case as positive in this particular situation; in fact, we would have the maximum of TP and no FN. Since we are considering only α and δ as relevant parameters in our GT function, we could claim to have achieved a greatest result. But, obviously this is not what we are looking for.

6.4 Some final remarks

In this section we try to draw some final consideration. In particular, in our opinion it is interesting to consider the trend of the four components of the confusion matrix, that we can observe in Fig. 6.8. In particular, we can note:

- TP and FN curves intersect in a particular point X_0, in which we can identify exactly half of the positives case (correct identification of the half of the elements in minority class).
- FP and TP curves both increasing if we reduce the imbalance of the training set (i.e. if we move from right to left along the x-axis in the graph). These two curves assume approximately the same slope up to point X_0 on the x-axis, at which they intersect. Instead, to the left of X_0, FP curve hangs much more than TP curve.

The previous 'graphics' observation related to the behaviour of the confusion matrix appear relevant in order to the evaluation of the prediction results. In fact, as just mentioned before, the dynamic of the slopes of the TP and FP curves (and consequently of the TN and FN) assume a different behaviour on the left of the point X_0. This fact could become very relevant in the decision whether to use or not use a more balanced training set. If we are interested to the correct classification of minority class (TP) we can increase the number of True positive if we use a more balanced training set. But this achievement lead to a cost that we pay with an increment of errors (FP) and with a reduction of TN.

It is interesting to note that if we use AuROC as performance indicator we obtain the maximum results with a balanced training set, because AuROC particularly rewards a high recall and therefore the high number of TP. But, if we are interested also in Precision we will instead be penalized by the increase in FP that derives from a training set more balanced. In this case the use of F1-score (or MCC) as performance indicators can represent a better choice.

Finally, in the previous case studies we found that also some gain functions set taking into account the needs of some stakeholders would indicate that a fully balanced training set is not always the best choice.

To conclude, we assert that the convenience in the use of a more balanced training set is highly dependent from the different slopes of the curve TP and TN and from the relative values that we attribute to the elementary four components of the confusion matrix.

We try to illustrate this evidence with some figures that report the variation of AuROC, F1-score and the GT function we used in the first use case. In particular, we calculate the variation of this performance indicators in function of the variation of TP and TN over the dataset composed of 1000 elements (of which only 40 positive). We can observe in Fig. 6.12 that AuROC is more sensible to the variation long the TP axis respect, for example, to the GT function (see Fig. 6.14). Also the dynamic of the F1-score indicator seems less sensible to the variation of TP. This evidence motivate the convenience to use a more balanced training set (that improve the number of TP) when we measure the performance using AuROC.

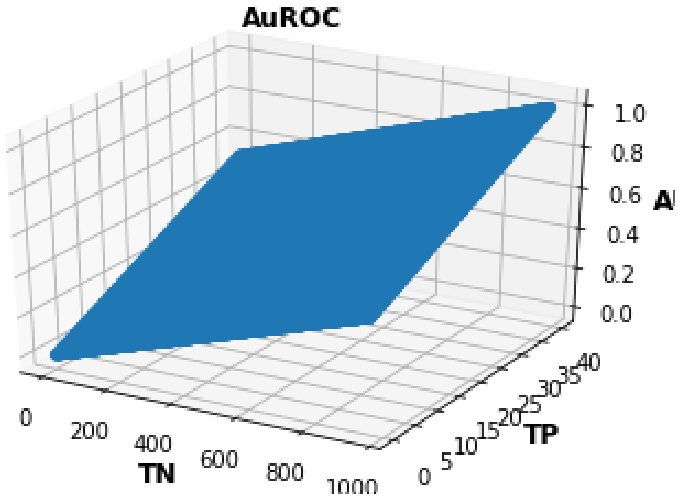


Figure 6.12: The figure show the dynamic of the AuROC in function of TP and TN.

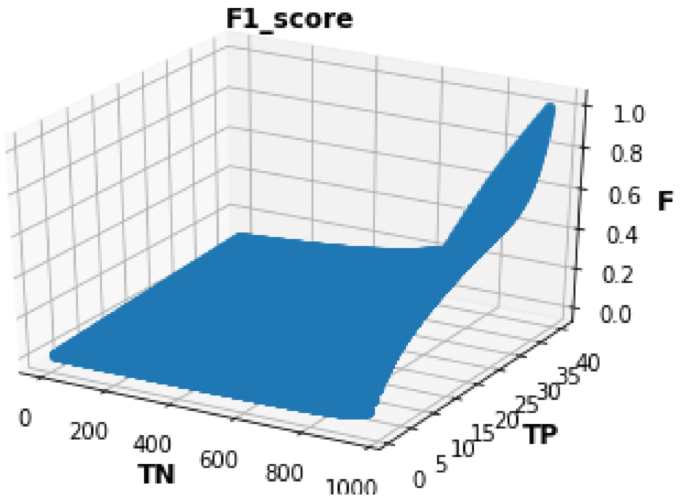


Figure 6.13: The figure show the dynamic of the F1-score in function of TP and TN.

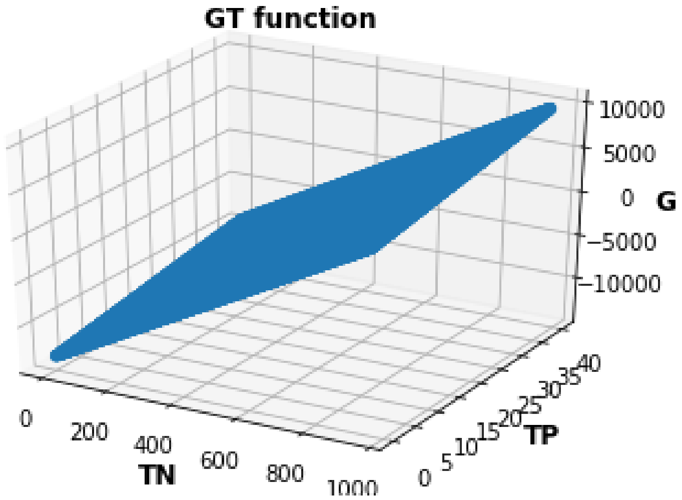


Figure 6.14: The figure show the dynamic of our GT in function of TP and TN. In this case the coefficient of GT function are $\alpha = 0$, $\beta = 10$, $\gamma = 10$, $\delta = 100$.

To summarize, we draw the following conclusions:

- The use of a fully balanced training set (which maximizes AuROC and Recall) does not represent the best approach in general, when we work in a highly imbalanced contest.
- We conjecture that there exists a well defined range in the training set imbalance in which we can determine the optimal configuration. More in details, with reference of Fig. 6.8, the optimal point will be included between $ratio = 1$ and $ratio = X_0$. Moreover, we found that the point X_0 depends on the test set imbalance.
- Performance evaluation task is highly dependent on specific needs. We hypothesize that a gain function (GT) customized to the needs of the stakeholders is probably the

best way to evaluate results in highly specialized contexts. The GT function have to take in account all the four components of the confusion matrix. With reference to the imbalance of the training set, the use of a customized GT function lead us to hypothesize greater gains using a slightly unbalanced training set (ratio between majority class and minority class between 3 and 5). Among the best known, F1-score and MCC are the two performance indicators that appear best suited to measure the results in such a scenario. In addition we found that F1-score and MCC show a very similar behaviour in a highly unbalanced scenario.

Chapter 7

Real applications and future works

In this chapter we put in practice our default prediction framework in order to applies it to some relevant real situation. The objective will be to better evaluate the effectiveness of our results obtained using very advanced ML techniques and data manipulation.

7.1 Comparison with External Ratings

In the previous discussion relating to explainability we have seen how some indicators especially developed to predict the vulnerabilities of companies are particularly important also in connection with our prediction obtained with ML techniques. In this section we compare our default prediction framework with two relevant of those ratings, namely the (1) Probability of default (PD), a score estimated by private banks and reported to the Central Bank in AnaCredit survey, and the (2) Cerved "rating", a score calculated by Cerved Group, probably the most authorative company that monitors and evaluates the balance sheet data of Italian companies.

In particular, in the two experiments we assume that we will predict a default if a firm has a probability of default greater than an established threshold (both for PD and ratings). Our prediction results show a very low performance demonstrating the difficulty of default prediction problem. In both cases we can see (Figure 7.1) that it is not possible to obtain an AuROC greater than 0.8. Contrast this with the AuROC score of 0.95 achieved by our approach using our Merged dataset (see Section 4.2.2).

In this single prediction explanation is important to notice how despite the fact that the company has the most important attributes, the probability of default, very low it succeed in predicting the company properly basing it on all the other features and showing how they influenced it.

The rich data that we use for our research, together with the careful data analysis, allow us to exceed, significantly, the standard external ratings. This appears a remarkable result as we have previously seen that the Probability of Default (PD) and Cerved ratings are two of the most relevant forecasting factors even using ML algorithms.

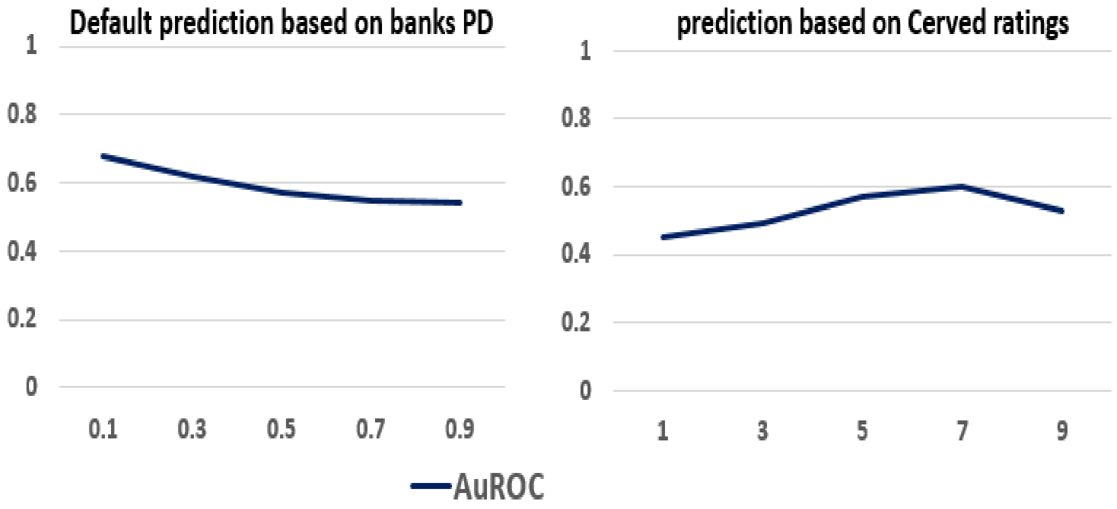


Figure 7.1: Adjusted default prediction based on bank's PD and Cerved ratings. PD is a number between 0 and 1 while the rating indicator ranges from 1 (best case) to 9 (worst case).

7.2 A practical application: Probability of Default for loan subgroups

We now see an application of our ML classifier in an applied problem faced by the Italian National Central Bank (Bank of Italy).

We compare our combined classifier (**COMB**) with a method commonly used to estimate the probability of one-year default by companies at aggregate level.

Consider a segmentation of all the companies (e.g., according to economic sector, geographical area, etc.). Often there is the need to estimate the probability of default (PD) of a loan in a given segment. A very simple approach, which is actually used in practice in Bank of Italy, is to simply take the ratio of the companies in the segment that went into default at year T+1 over all the companies that were not in default in year T. We use this method as a baseline.

We now consider a second approach based on our classifier, which we call **COMB**. We estimate the PD by considering the amount of companies in the segment that are expected (using the **COMB** classifier) to go into default at year T+1 compared to the total loans existing for the segment at the time T.

We use two different segmentation. A coarse one, in which the segments are defined by the economic sector (e.g. mineral extraction, manufacturing) , and a finer one, which is defined by the combination of the economic sector and the geographic area, as defined by a value similar to the company's zip code (CAP).

In Table 7.1 we compare the two approaches for estimating the PD. As expected, in both segmentation the classifier-based approach is a winner, with the improvement being larger for the finer segmentation. In many cases the two approaches give the same result, typically because in these cases there are no companies that fail (PD equals to 0).

Coarse segmentation		
	Baseline	**COMB**
Mean error	0.11	0.048
Var error	0.056	0.016
Superiority percentage	25.1%	45.6%

Fine segmentation		
	Baseline	**COMB**
Mean error	0.088	0.036
Var error	0.06	0.025
Superiority percentage	6.1%	19.5%

Table 7.1: Comparison of the standard approach to estimate PD with the classifier-based one. "Mean error" is the average error between the predicted PD value and the real one. "Var error" is the variance of the error. "Superiority percentage" is the percentage of segments in which the predictor is better than the other; in the remaining ones we have the same performance.

7.3 A Current Application: an assessment for Covid credit guarantees

The Covid-19 pandemic had an unprecedented, dramatic impact on the corporate sector across the globe, which was very swiftly met by an equally unprecedented response by Governments and Central banks. Despite differences in the design, size and timing of the interventions, virtually all Governments offered state guarantees on new loans to ensure that bank credit continued to flow speedily and cheaply to the economy. In Italy, the Government acted by significantly strengthening the existing guarantee program from the Central Guarantee Fund ("Fondo Centrale di Garanzia", FCG) and complementing it with a new one under the supervision of the export credit agency (SACE). Between the end of February 2020 and the end of September 2020 the flow of granted loans guaranteed by Italian Government was about 73 billion euro; these loans represented by far the main driver of credit in the period.

In this section, we provide an assessment of the potential cost (and potential savings) for public finances that guarantee these loans.

The crucial point is related with the need to balance the risk of losing public money against the risk of not financing a company that would be able to survive. For a company destined to fail, the state may decide to not provide guarantees. Thus, we compare the savings between using the probability of default index (PD) and our ML approach.

Assume that we use the PD approach, that is, we use the probability of default (PD) attributed by banks to their own borrowers. See Table 7.2. Using different PD thresholds, we can try to predict firms defaults and estimate public guarantees that will be lost after only one year. The third column shows how much we expect to save and how many guarantees we deny to a subset of the economic system that could benefit from them. The rightmost column of the shows the amount of guarantees that we could have predicted as lost one year in advance using PDs.

Consider instead using our ML approach (e.g., CatBoost). By varying the ratio of positive and negative instances in the training set, we can vary the performance measures of the

PD	Precision	Amount of lost guarantees predicted (millions)	Amount of lost guarantees (millions)
0.001	0.05	70196	3510
0.1	0.15	2318	348
0.3	0.27	918	248
0.5	0.40	661	264
0.7	0.47	543	255
0.9	0.45	437	197

Table 7.2: Predictions of the loss of State guarantees after one year following company bankruptcies based on the use of the probability of default (PD) used by Italian banks. These PD values are those used by banks for supervisory purposes and are reported to the Bank of Italy as part of the AnaCredit survey.

unbalance training set (%)	Precision	Amount of lost guarantees predicted (millions)	Amount of lost guarantees (millions)
100	0.18	19070	3433
14	0.46	7670	3528
8	0.53	1811	960
6	0.60	1527	916
4	0.67	1220	817
2	0.75	814	611

Table 7.3: Predictions of the loss of State guarantees after one year following company bankruptcies based on the use of the predictions of Catboost using our credit dataset.

classifiers, in particular the precision. Replicating the process described above for PD, we obtain Table 7.3. The results show that we can significantly increase the ability to predict expected losses. In particular, by performing a prediction using a highly unbalanced training set (and then choosing to work with high precision in prediction), we could estimate a saving of over 600 million euros for public finance. In contrast, the the cost of this selection would be not providing only 200 euros million to companies that would not have gone bankrupt. Obviously, even in a scenario like this, the ability to understand and explain decisions plays a fundamental role.

7.4 Future works

According to the results showed in this thesis, it can be said that some improvements from the tool that are still used have been made. The financial system may starts taking into consideration to implement or integrate tree based machine learning models into their processes, together with an appropriate explainability tool. What is possible to say is that the problem that is addressed here is and will be hard maybe until the end of time. Anyway here it is showed that new algorithms and models with new data merged together may be a right path to follow. May not be a long way since Deep neural networks will outperform tree base algorithms, maybe it is just a matter of time and something to deep dive into. Another point that will not deserve to be left on the side is explainability. In many areas and sectors, explainability will be the reason why the adoption of machine learning will really start. Even though it seems a field with a much slower growth and innovation rate, it will be fundamental to build what is called Auto ML with an high degree of trust. As well as while talking about the future of DNN for tabular data in terms of performance, explainability with model-agnostic algorithms must expand to them with the same degree of confidence that has been achieved with ML models.

Chapter 8

Conclusion

Business-failure prediction is a very important topic of study for economic analysis and the regular functioning of the financial system. Moreover, the importance of this issue has greatly increased following the recent financial crisis. Furthermore, we can certainly consider that the current global crisis caused by Covid-19 will lead to a significant increase in business failures, adding further relevance to the ability in prediction of firms bankruptcy. Another fundamental issue concerns the possibility to explain the results obtained in bankruptcy prediction. The first studies relating to the prediction of default date back to over 50 years ago; more recently there have been many studies that have tried to predict the failure of companies using various Machine learning (ML) techniques.

In our project, we try to predict the banking default of Italian companies, using Machine Learning and other well-known statistical techniques. We focused our attention on bankruptcy and bank default, using three different source of data:

- credit information from the Italian Central Credit Register (CCR),
- public balance sheet data and, for the first time,
- credit data from ECB AnaCredit survey.

We analyzed a very large amount of credit data containing information about almost all the loans of all the Italian companies. Our first findings is that, both in the case of bankruptcy prediction and bank default prediction, ML approaches are able to outperform significantly simpler statistical approaches.

In fact, our results confirm the best performance of ML classifier respect to other well-known statistical methods (see [7]). In addition we show that some recent types of boosting classifiers obtain the best results.

Furthermore, in this project we explored the differences and links between corporate failures and bank defaults. We focus our analysis both on bankruptcy, a well-known issue in literature, and on the bank default, which in many cases anticipates the failure of a company. At the same time it represents an important sign of vulnerability as typically a company in bank default is unable to repay its debts. In our knowledge it is the first time that both these status of difficulty of companies are considered in the same study.

We reach a remarkable result in prediction also with reference to bank default; this provides the basis for the use of bank default prediction also in the search for bankruptcy, given the connections between the two conditions of vulnerability of companies which we highlighted in the first chapter.

We use credit data in combination with balance sheet data demonstrating that this combination of data can lead to better performance in prediction. In fact, using information on past loan data is crucial, but the additional use of balance-sheet data can improve classification even further. We also show that using loan data in the prediction of bankruptcy (where, typically, only balance-sheet data are being used) can play an important role.

The forecast performance obtained both for bankruptcy and for bank default prediction are very similar, but a relevant point seems to be that balance sheet data are more suitable for predicting bankruptcies; otherwise the loan data helps to predict bank defaults much better. Moreover, we corroborate this conjecture also in terms of expainability of the prediction results. Another interesting point concerns the depth of the dataset necessary for the predictions: we show that for credit data it is not necessary to go far back in time to have the best predictive performance, but that one year of past data is enough. Also for balance sheet data we show that a very long time series do not improve in a significant way the final performances.

In fact, we show in Section 4.3 that the use of a lot of backward data (beyond a certain limit that we have identified with 4 quarters) does not significantly improve the predictions. This is in line with the explainability analysis that show us that the features relating to the most recent dates drive the prediction and that the features that take into account the variations with respect to past reference dates do not work very well.

In addition we use, for the first time, also information data from ECB AnaCredit survey, recently started under the coordination of ECB, in order to improve bank default prediction.

In our experiments, we try to exploit the predictive capacity of our credit information by using BORUTA, a feature selection technique that seems to work very well in our case.

Our approach allowed us to obtain very remarkable results, with an AuROC of 0.97 for bankruptcy prediction using credit data in combination with balance sheet data and a similar performance (AuROC=0.95) for bank default prediction. These results represents truly remarkable performances also in comparison to the best results found in the literature on the subject.

Moreover, a relevant point of our work concerns the attempt to explain default predictions; this issue is indeed very relevant for the practical use of forecasting techniques. To this end we used Shap, a modern method to extract the importance of features in the forecast, showing a robust dependence of our predictions on a series of information that have a clear meaning and an important economic significance. For example we have highlighted a significant explanatory power of the geographical location of a company. Moreover, as easily understood, our explainability assessment shows clearly that some very relevant well-known indicators like Probability of default (PD) assessed by the banking system and Cerved *rating* (based on balance sheet data) play a significant role both in bankruptcy and in bank default prediction. So, we can try to use this results in order to better understand our predictions. Moreover, it is interesting to note that a comparison exercise with the default prediction based only on PD used by banks shows that predictions with machine learning provide a very significant gain in performance (Chapter 7.1).

In addition we show (see again Chapter 7.1) that ML techniques outperform also predictions based only on Cerved ratings, that are well-known scores related to firms vulnerability.

A further outcome of our work is the attempt to go beyond the academic and theoretical study to put into practice the results that can be obtained in default prediction by comparing with the techniques actually used on this field.

In this regard, a significant part of the work is specifically dedicated in developing a framework to maximize the performance of predictions, taking into account the specific highly unbalanced scenario that characterize our reference contest.

We conjecture that the use a perfectly balanced training set (via SMOTE o via under sampling) does not represent in many cases the best choice. We perform an in-depth analysis and we propose a total gain function in order to establish a system for evaluating the performance of individual predictions.

With this regard, we draw the following conclusions:

- We show that the use of a fully balanced training set (which maximizes AuROC and Recall) does not represent the best approach in general, when we work in a highly imbalanced contest.
- We conjecture that exist a well defined range in the training set imbalance in which we can determine the optimal configuration.
- Moreover, we acknowledge that performance evaluation task is highly dependent on the specific metrics that we use but also on the specific business needs. We hypothesize that a gain function customized to the needs of the stakeholders is probably the best way to evaluate results in highly specialized contexts.

Finally, our analysis leads further to the following other considerations: i) in our specific scenario we can obtain the greater gains using a slightly unbalanced training set (ratio between majority class and minority class between 3 and 5); ii) among the best known performance indicators, F1-score and MCC are the two metrics that appear best suited to measure the results in such a scenario; iii) in some specific scenario (like for example bank risk assessment) the use of a customized gain function could be the preferable choice in order to evaluate the performances.

Nevertheless, default prediction remains an extremely hard problem. Yet, even slight improvement in the performance, can lead to savings of multiple hundreds of thousands of euros for the banking system. The ability to accurately predict bankruptcies can also lead to savings for public finances, as we try to highlight using the specific case of the recent "Covid" measures to help the economy. But, especially in cases like this, it is even more important to be able to explain the predictions obtained and the consequent choices.

The last point we would mention is related with the data. We assume that part of the difficulty of the task lies also in the incomplete adequacy of the data. Balance sheet data are available only annually and with a long delay; credit data are available with a higher frequency but can sometimes hide the real situation of companies. Moreover, we have shown that data that is too old is no longer very useful. Exploring other sources of information would certainly be a promising objective to pursue in the future.

Bibliography

[1] Methods and sources: Methodological notes, 2018. Available on the website of the Banca d'Italia at `https://www.bancaditalia.it/pubblicazioni/condizioni-rischiosita/en_STACORIS_note-met.pdf?language_id=1`.

[2] S. Albanesi and D. F. Vamossy. Predicting consumer default: A deep learning approach. 2019.

[3] T. Aliaj, A. Anagnostopoulos, and S. Piersanti. Firms default prediction with machine learning. In V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, S. Pascolutti, and G. Ponti, editors, *Mining Data for Financial Applications*, pages 47–59, Cham, 2020. Springer International Publishing.

[4] E. Altman. Predicting financial distress of companies: Revisiting the z-score and zeta. *Handbook of Research Methods and Applications in Empirical Finance*, 5, 2000.

[5] M. Andini, M. Boldrini, E. Ciani, G. de Blasio, A. D'Ignazio, and A. Paladini. Machine learning in the service of policy targeting: the case of public credit guarantees. (1206), 2019.

[6] A. Atiya. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *Neural Networks, IEEE Transactions on*, 12:929 – 935, 2001.

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

[8] J. Begley, J. Ming, and S. Watts. Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Review of Accounting Studies*, 1:267–284, 1996.

[9] J. Boritz, D. Kennedy, and A. d. M. e. Albuquerque. Predicting corporate failure using a neural network approach. *Intelligent Systems in Accounting, Finance and Management*, 4(2):95–111, 1995.

[10] A. H. O. K. S. F. L. Branka Hadji Misheva, Joerg Osterrieder. Explainable ai in credit risk management. 2021.

[11] C. Chakraborty and A. Joseph. Machine learning at central banks (september 1, 2017). Bank of England Working Paper No. 674, 2017.

[12] M.-Y. Chen. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, 62(12):4514 – 4524, 2011.

[13] S. Cho, H. Hong, and B.-C. Ha. A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the mahalanobis distance: For bankruptcy prediction. *Expert Systems with Applications*, 37(4):3482 – 3488, 2010.

[14] B. Erdogan. Prediction of bankruptcy using support vector machines: An application to bank bankruptcy. *Journal of Statistical Computation and Simulation - J STAT COMPUT SIM*, 83:1–13, 2012.

[15] L. et al. From local explanations to global understanding with explainable ai for trees. *Nat Mach Intell*, 2:56–67, 2020.

[16] L. P. et al. Catboost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems.*, page 6637–6647, 2018.

[17] A. Z. Evgeny Lyandres. Investment opportunities and bankruptcy prediction. *Journal of Financial Markets*, 6 (2015), 2013.

[18] E. Fernandez and I. Olmeda. Bankruptcy prediction with artificial neural networks. In *In Proc. of the 2018 2nd International Conference on Inventive Systems and Control (ICISC 2018)*, volume 930, pages 1142–1146, 1995.

[19] A. Fuster, M. Plosser, P. Schnabl, and J. Vickery. The Role of Technology in Mortgage Lending. *The Review of Financial Studies*, 32(5):1854–1899, 04 2019.

[20] A. Gepp and K. Kumar. Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science*, 54:396–404, 2015.

[21] F. F. Javier Garcia. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 6 (2015):1437–1480, 2015.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS*, 2017.

[23] P. R. Kumar and V. Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review. *European Journal of Operational Research*, 180(1):1 – 28, 2007.

[24] M. B. Kursa and W. R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(2), 2010.

[25] H. Kvamme, N. Sellereite, K. Aas, and S. Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018.

[26] T. Le, M. Y. Lee, J. R. Park, and S. W. Baik. Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset. *Symmetry*, 10(4), 2018.

[27] S. Lee and W. S. Choi. A multi-industry bankruptcy prediction model using back-propagation neural network and multivariate discriminant analysis. *Expert Systems with Applications*, 40(8):2941 – 2946, 2013.

[28] W.-C. Lee. Genetic programming decision tree for bankruptcy prediction. In *9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press, 2006.

[29] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai. Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man, and Cybernetics - TSMC*, 42:421–436, 2012.

[30] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4768–4777, 2017.

[31] E. Martinelli, A. de Carvalho, S. Rezende, and A. Matias. Rules extractions from banks' bankrupt data using connectionist and symbolic learning algorithms. *Proc. Computational Finance Conf*, 1999.

[32] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. arXiv:1706.07269v3 [cs.AI] 15 Aug 2018, 2018.

[33] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[34] M. Moscatelli, F. Parlapiano, S. Narizzano, and G. Viggiano. Corporate default forecasting with machine learning. *Expert Systems with Applications*, 161:113567, 2020.

[35] L. Nanni and A. Lumini. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Syst. Appl.*, 36(2):3028–3033, 2009.

[36] A. Narvekar and D. Guha. Bankruptcy prediction using machine learning and an application to the case of the covid-19 recession. *Data Science in Finance and Economics*, 2021.

[37] M. Odom and R. Sharda. A neural network model for bankruptcy prediction. In *In Proc. of the 1990 IJCNN International Joint Conference on Neural Networks*, volume 2, pages 163 – 168 vol.2, 1990.

[38] J. A. Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18-1, 1980.

[39] A. S. Paolo Giudici, Branka Hadji-Misheva. Network based credit risk models. *Quality Engineering*, 32-2:199–211, 2020.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[41] A. A. Ravid Shwartz-Ziv. Tabular data: Deep learning is not all you need. *8th ICML Workshop on Automated Machine Learning (2021)*, 2021.

[42] M. E.-A. Sabri Boughorbel, Fethi Jarray. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *Expert Systems with Applications*, 2017.

[43] S. Sarojini Devi and Y. Radhika. A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing*, 8:133–139, 2018.

[44] R. Sharma, C. Schommer, and N. Vivarelli. Building up explainability in multi-layer perceptrons for credit risk modeling. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 761–762, 2020.

[45] I. K. Sophia Daskalaki and N. Avouris. Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20:5:381–417, 2006.

[46] R. A. J. Sudheer Chava. Bankruptcy prediction with industry effects. *Journal of Financial Markets*, 6 (2015), 2004.

[47] G. Wang, J. Ma, and S. Yang. An improved boosting based on feature selection for corporate bankruptcy prediction. *Expert Systems with Applications*, 41(5):2353 – 2361, 2014.

[48] N. Wang. Bankruptcy prediction using machine learning. *Journal of Mathematical Finance*, 07:908–918, 2017.

[49] L. Zhou and H. Wang. Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 10:1519–1525, 2012.

Acknowledgements

A very special thanks to Professor Aris Anagnostopoulos that followed me along this long journey, also in my hardest moment. I'm sure I've improved, thanks also to its help; if he wasn't there I wouldn't be here today. I am proud to have shared a large part of this work with Valerio Evangelisti, which helped me a lot.

Stefano Piersanti, Rome, April 2022