**Università degli Studi di Padova**

FACOLTÀ DI MATHEMATICAL MODELLING FOR ENGINEERING
Corso di Laurea in Financial Engineering

Tesi di laurea magistrale

# Anticipation of default: implementation of a credit scoring model and integration with macroeconomic factors

Candidato:
**Davide Ragnoli**
**Matricola 2091431**

Relatore:
**Prof. Martino Grasselli**

**Anno Accademico 2023-2024**

**Abstract**

This thesis explores the anticipation of credit default through the development and implementation of a credit model. Anchored in the context of the Italian banking sector, it reviews credit risk regulations, such as Basel III and IFRS 9, and their implications for risk management strategies. Leveraging advanced machine learning techniques, including gradient boosting and random forests, the study aims to design an early warning system that predicts defaults with greater precision. Furthermore, the model incorporates key macroeconomic variables, such as inflation, house prices, and interest rate fluctuations, to assess their impact on credit risk.

The purpose of the study is to determine whether advanced ML techniques can significantly improve the credit risk monitoring process by intercepting the majority of positions that exhibit credit difficulties and whether the inclusion of macroeconomic features leads to different results.

# Contents

# Chapter 1

# Banking sector and risk management

## 1.1 The Italian banking system

The banking system is one of the most fundamental pillars in any modern economy that acts as a liaison between people with access to capital and those who do not. As reported in [Giordano and Lopes 2008], banks provide vital services through deposit-taking from both individuals and institutions and play a very significant role in the process of advancing money for loans and investing in various sectors for the cause of economic advancement. They, in return, facilitate consumption, business expansion, and overall development of the economy. Although it might be hard to precisely define what comprises a "banking system," one can say that it is generally composed of the network of financial institutions, starting from a small, local bank and extending to large international banks, that provide financial services in the form of loans, deposits, investment opportunities, and money management services [Ongena and Smith 2001].

The Italian banking system has undergone major changes, especially in the last decades, transforming from an essentially state-controlled model to

a privatized and more competitive one. For a very long period, in fact, up until the 1990s, the Italian banking sector was dominated by government ownership, leaving very little room for private competition, and was mostly oriented toward the domestic market [Giordano and Lopes 2008]. By interfering unduly, the state kept this sector on a tight leash, often with inefficient performance and retarded growth.

The turning point took place in the year 1990, with the so-called Amato-Carli Law, which traced a new path into the Italian banking world. According to the Amato-Carli Law [Giordano and Lopes 2008], public banks could change their status to joint-stock companies , therefore taking private investment and opening space for competition in this field. This move to liberalization was sealed more appropriately by the Consolidated Banking Act of 1993, which laid the legal framework for a more market-oriented system. These reforms opened a new perspective that allowed Italian banks to become modern, expand, and at the same time be more competitive both internally and on an international level [Nikolopoulos and Tsalas 2017].

Despite such advances, over the years, the Italian banking system has encountered numerous problems and the credit aspects have always been at the forefront, as reported in [Barbagallo 2018]. Of course, credit is the very heart of banking activity, and yet it has proved one of the most troublesome areas. Italian banks have had to put up with a really huge amount of NPLs that are loans which are at risk of not being repaid, usually because of the poor financial condition of a borrower. By 2015, according to [Barbagallo 2018], NPLs increased to a very high level of €200 billion in Italy due to serious concerns over the stability of the banking system. Banks that are heavily laden with high levels of NPLs face liquidity challenges; thus their capacity for the issuance of new loans is greatly limited in support of economic growth [Nikolopoulos and Tsalas 2017].

In turn, Italian banks due to the crisis (each on the back of regula-

tory measures) made strong efforts to decrease the volume of NPLs in their portfolios [Nikolopoulos and Tsalas 2017]. Italian banks, once the IFRS9 standard and the calendar provisioning (a tough regulatory measure that compelled banks to gradually write down the value of the non-performing loans) had taken place, managed to bring the NPL ratio down to 7.6% by 2017 [Barbagallo 2018]. Indeed, this was a remarkable improvement that reflected positively in terms of purification of balance sheets and enhancement of asset quality in general. Non-performing loans, however, remain a threat to the system, and Italian banks are supposed to continue showing vigilance while managing these risks.

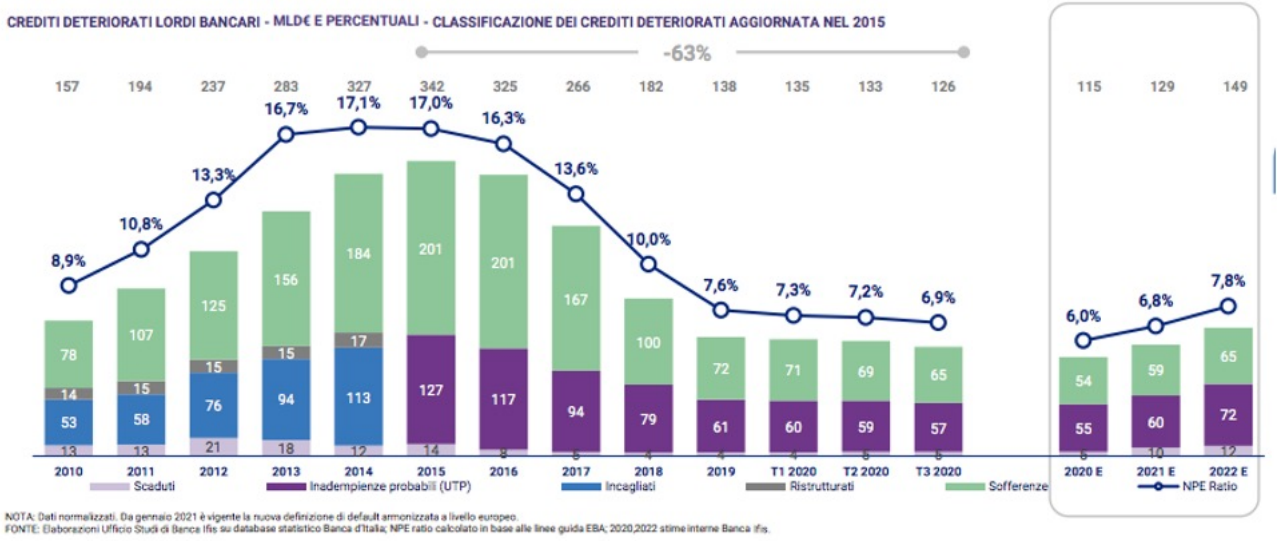


**Figure 1.1:** Distribution NPL from 2010 to 2020, from **[barbagallo2018sistema]**

The second challenge the Italian banking systems are facing is that their profitability is relatively low as compared to the European average. Among others, this may be due to the fact that Italy has very poor economic growth, and also because of the cost of risk management, which refers to the expenses taken by the banks with regard to addressing potential losses resulting from bad loans. Although the Italian banks showed partial improvement in profitability during 2017 (where ROE had swelled up to 4% after remaining

negative the previous year, as reported in [Barbagallo 2018]) profitability remains an issue. This means that Italian banks, if they want to remain competitive with European banks, will have to continue looking for ways to increase their revenues and reduce costs.

Aside from internal issues, the Italian banking industry is also faced with external competition from fintech companies ([Barbagallo 2018]). Short for financial technology, fintech refers to the firms using innovative technology to offer financial services in an efficient and user-friendly manner. Due to this fact, the operations costs for a fintech firm are usually lower than those of a traditional bank, hence they can charge lower prices for their services [Ongena and Smith 2001]. Of course, all this puts a lot of pressure on traditional banks to reassess their business models, enhance their technological offerings and innovatively invest in competitiveness. In sum, it's changing the way banks do business, becoming far more agile and responsive to emerging consumer expectations [Lamarre, Smaje, and Zemmel 2023].

The distinctive peculiarity of the Italian banking system is the variety of its structure: big internationally oriented banks coexist with small local ones in Italy. The last group includes cooperative credit banks-common purely community-oriented financial institutions of first importance for SMEs [Altman and Sabato 2007]. SMEs are fundamental to the Italian economy and local banks, like BCCs, contribute to enterprise growth by providing finance. This close link between banks and local enterprises is a peculiar feature of the Italian financial system.

On the outlook, some positive aspects are to be noted in the Italian banking system: indeed, the Italian non-performing loans are falling, and bank capitalization is on the rise. One of the relevant indicators of such progress is the Common Equity Tier 1 ratio, which determines the degree of a bank's financial strength through its core capital equity in relation to its total risk-weighted assets. As written in [Barbagallo 2018], during the

two years from 2015 to 2017, the CET1 ratio of banks in Italy went up from 12.3% to 13.8%, reflecting, therefore, a strengthening of their financial position. With these capital reserves being built up, banks become resistant to various financial shocks and generally, this makes the entire system more stable. More significant capitalization added to the cleaning up of NPLs had put the Italian banking system in a better place.

There are still challenges, however. Profitability remains an issue, and the return enjoyed by many Italian banks is below the levels considered desirable. With that in mind, the only way any bank could address this would be to make sure that they diversified their sources of revenue.

This might involve asset management and other non-traditional services with regards to finances that gave an avenue for creating revenue outside of the traditional model of lending [Lamarre, Smaje, and Zemmel 2023]. Additionally, Italian banks will also need to invest heavily in technology as financial services are becoming increasingly digital. This would be a way for them not only to compete with Fintech companies, but also for them to meet consumers whose expectations continue to evolve: in fact they demand speedier, more accessible and user-friendly banking services [Barbagallo 2018].

Regulatory reform has also greatly affected the Italian banking system. Global regulations, such as Basel III (which details minimum requirements of capital for banks, [Vousinas 2015]) and the Bank Recovery and Resolution Directive (dictating how to deal with failing banks) have placed stricter controls related to liquidity and capital reserves. It is these regulations that have strengthened the Italian banks but also increased their cost of operation in compliance particularly.

Going forward, the Italian banking system faces a set of critical issues: in a low interest rate environment with economic growth at a snail's pace, banks will have to improve their operational efficiency and cut costs in order to sustain profitability. Besides, they will be expected to further pursue the

financing of the real economy, especially credit to enterprises, with great emphasis on SMEs that continue to represent the backbone of the Italian economy. With ongoing changes facing Italy's banking sector, how it could manage profitability, innovation, and stability would be its future.

## 1.2 Credit Risk regulation

The two major regulations that reshape the regulatory and accounting framework for financial institutions globally are Basel III and IFRS 9. While separated by scope and objectives, their purpose is to enhance the stability of the global financial system in view of past financial crises [Altman and Sabato 2007].

As reported in [Beerbaum and Ahmad 2015], "IFRS 9 is an accounting standard that introduces an expected loss model for the valuation of financial assets". Unlike the incurred loss approach applied previously, IFRS 9 requires institutions to account not only for losses already incurred but also for expected ones, in order to anticipate the recognition of losses in the financial statements [Löppönen 2022]. The model is divided into three stages, each one of them corresponding to a different level of deterioration of the credit risk of a financial asset. In the first stage, an asset is considered to have low credit risk. Expected credit losses should be measured in the first instance over the ensuing twelve months regardless of whether a loss has occurred. If credit risk has increased significantly since initial recognition, the asset moves into the second stage where expected losses are measured over an asset's lifetime ["Basel committee on banking supervision" 2011]. If finally, in the third stage, objective evidence of impairment of credit exists, losses are fully recognized and interest is calculated on the net-carrying amount of the asset, net of loss allowances.

The IFRS 9 stages are divided, as cited in [Beerbaum and Ahmad 2015], as follows:

- Stage 1: Assets are considered to have low credit risk. Expected credit losses are measured over the next 12 months, even when there has been no loss as yet.

- Stage 2: Where there has been a significant increase in credit risk, the credit risk is not as high in Stage 2. Now, the expected credit losses are measured during the lifetime of the asset.

- Stage 3: If there is any objective evidence of credit impairment, the full loss is recognized and interest is calculated on the net carrying value of the asset net of loss provisions.

It is in this monitoring of credit risk, then the exercise of judgment in terms of updating the loss estimates, where the complexity of IFRS 9 is deemed to be [Beerbaum and Ahmad 2015]. In this respect, the development of new systems and processes, if not just enhancements to existing ones, would be required from the financial institution. A financial institution needs to gather and process immense volumes of information in order to express the expected loss model in a comprehensive way. This accounting change has therefore significantly altered the way banks portray their profitability and their capital requirements by increasing volatility in earnings and by raising their credit loss provisions.

Basel III, by contrast, is a set of rules and regulations aimed at strengthening the international banking system from any future financial crisis that might occur post-2007-2008 [Vousinas 2015]. It mainly focuses on strengthening the quality and quantity of capital by banks, besides proposing two new liquidity standards. The two major ratios proposed herein are the Liquidity Coverage Ratio and the Net Stable Funding Ratio [Löppönen 2022]. It is in contrast to the LCR, which requires banks to hold an adequate level of high-quality liquid assets to enable them to survive liquidity stress in the short run and meet expected cash outflows at least for a 30-day period. Whereas

the NSFR tackles the long term: it promotes those banks that have a more stable funding structure by decreasing their dependence on short-term and unstable funding sources.

Furthermore, under the liquidity requirements, Basel III has stipulated the leverage ratio with an aim to prevent banks from excessive leverage. In particular, the leverage ratio that refers to Tier 1 capital/total assets is supposed to diminish risks, since a high level of leverage reflects a danger of loss rise after the crisis period ["Basel committee on banking supervision" 2011].

Due to these regulations, the financial entities have been hugely adapting their ways of handling risk management and accounting. On one hand, IFRS 9 made it imperative to devise models for expected credit losses, which becomes a far more subjective evaluation and considerably enhances operational complication [Beerbaum and Ahmad 2015]. On the other hand, Basel III comes with greater emphasis on capital buffers and liquidity management at higher levels (such aspects have influenced banks directly regarding their financial strategies and costs [Vousinas 2015]).

Both IFRS 9 and Basel III radically changed the way risk management of financial assets and accounting was done in banks. This has been put in place so that in the future, financial crises will be avoided and the banking system will be resilient to economic shocks, enhancing transparency and overall stability in the sector.

## 1.3 Credit Risk management

Credit risk management is very important in the financial world because it puts concentration on the protection of institutions such as banks, credit unions, and lending agencies against possible financial losses that emerge when debtors fail to fulfill their debt obligations.

This risk ensues from every kind of lending activity where the possibility

of borrower default is huge. In this respect, the credit risk management effectively ranges from the identification and measurement of these risks, along with their mitigation, in order to ensure the institution's long-term financial stability and integrity [Altman and Sabato 2007].

Credit risk demands a mix of qualitative judgments blended with quantitative models when assessing the creditworthiness of borrowers. Creditworthiness, in basic terms, refers to the probability that a borrower is able to pay off his debt obligations [Brown and Moles 2014]. Institutions use a range of methodologies for estimating this probability, more commonly referred to as default risk. The methodology can vary significantly depending on the type of borrower and the nature of the loan, including how much is borrowed, what the terms of the loan are, and what the loan will be used for. The ultimate goal is thus to arrive at the risk associated with lending to a particular borrower, identify those warning signals, and mitigate the risk as far as possible.

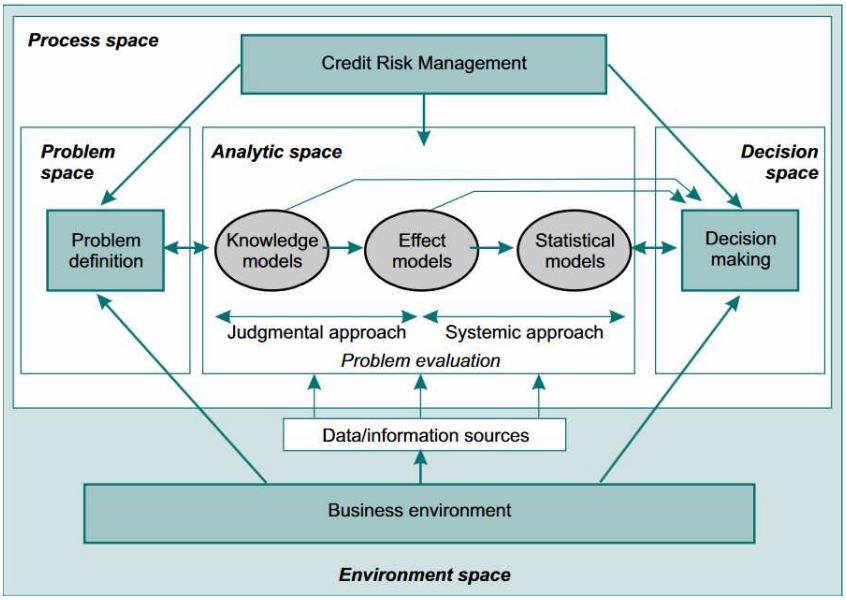


**Figure 1.2:** Credit Risk Management process, from [Brown and Moles 2014]

### 1.3.1 Various Credit Risk Assessment Approaches

There are several ways in which financial institutions assess credit risks, each with their own advantages and disadvantages. For instance, expert systems apply human experience and judgment. In this regard, analysts take into consideration several qualitative factors including the borrower's character, his or her ability to repay, financial capital available with the borrower, the collateral to back up the loan, general economic conditions prevailing and whether an applicant meets all legal requirements [Brown and Moles 2014].

Such multi-dimensional analysis provides the analyst with an overall picture of the borrower's financial position.

The other widely used system is the credit-scoring system, in which credit applicants are assigned a numerical score based on some preselected factors [Brown and Moles 2014]. The score expresses the perceived creditworthiness of the borrower, derived from an analysis of his or her financial history and market position. Credit scoring is particularly useful in evaluating credit applications by individual borrowers, as this provides a formal and objective estimate of the credit risk involved. These scoring models typically consider income levels, credit history, payment habits, and employment status to provide more data-oriented means of stratification.

By contrast, market-based models rely on real-time data from the market prices of financial instruments, particularly stocks and bonds. These models work on the principle that market prices encapsulate investor expectations for a company's future performance, including the associated risk, which encompasses credit risk. For example, if a credit spread on a corporate bond is perceived to widen relative to a risk-free government bond, that would imply an increase in the perceived credit risk of that corporate borrower [Bernanke 1993]. Market-based models provide valuable information about current market conditions and therefore enable an institution to revise its estimates of risks.

### 1.3.2 Economic Cycles and Their Impact on Credit Risk

The economic environment has a huge impact on credit risk, mainly because the overall economic cycles, marked by expansion and contraction, directly affect the borrowers' financial health. For example, in an economic boom, companies tend to show better earnings and cash flows, increasing their likelihood of meeting obligations. Because of this, financial firms tend to underestimate the risks in lending during these good times.

In bad economic times, business profitability will not be easy, and the rate of default will be higher. Such a turn can overestimate credit risk in the eyes of financial institutions because of how it will influence lending strategy and overall portfolio management.

In tracking these economic cycles, financial institutions' needed metrics include Probability of Default (PD) and Loss Given Default (LGD) [Brown and Moles 2014]. PD gives the probability of the occurrence of default within a certain period and thanks to this metric lenders can identify the potential risk that could be associated with certain borrowers.

On the other hand, LGD estimates the expected loss a lender may face in case a borrower defaults. This also considers various issues that relate to the value of collaterals as well as the general financial health of the borrowers. These together form a complete framework upon which to base and revise one's risk assessments in accord with shifting economic circumstances.

Effective credit risk management relies on close monitoring of some important credit events that may signal a borrower's financial distress. This includes missed payments, filings for bankruptcy, debt restructuring, credit rating downgrades, and the breach of loan covenants. These would be warning signals to show that a borrower is going into a problem area regarding his obligation.

In addition to monitoring such events, institutions focus on key metrics such as Loss Given Default (LGD), a measure of how much is at risk in

case of default. Such a measure will be determined, among other things, by the value of the collateral and the general financial condition of the borrower [Brown and Moles 2014]. The other key metric is Exposure at Default (EAD), defined as the amount the lender is exposed to when the occurrence of default takes place. This close monitoring enables them to be more aware of their risk exposures and make truly informed decisions about the extent of their lending.

### 1.3.3 Diversification: A Key to Credit Risk Mitigation

As written in [Brown and Moles 2014], diversification is the spreading of investments or loans over a wide range and it's the main approach to the mitigation of credit risk, as it reduces the overall level of portfolio risk. The process of spreading credit exposures across a huge number of borrowers, sectors, and geographic regions significantly reduces individual defaults. This is due to the fact that the impact of defaults on portfolio performance is minimized. For instance, when one borrower defaults, the impact of such a default can be cushioned by the presence of other performing loans, which in essence protects the financial position of the financial institution.

The diversified credit portfolio is usually less volatile due to the fact that the performances of different loans are usually uncorrelated individually and this will then provide an averaging effect whereby good performers may balance the losses of others. Besides, diversification contributes to the predictability of the losses.

The law of large numbers states that the larger the number of independent loans, the more the averages of the outcomes stabilize, therefore allowing institutions to anticipate their losses more appropriately and make preparations.

However, while diversification is an excellent tool in credit risk management, it isn't a panacea. Successful diversification requires active manage-

ment of the portfolio. In other words, one should constantly monitor loans, look out for any possible concentration of risks, and rebalance one's portfolio as market conditions or borrower performance changes.

### 1.3.4 Comparing Statistical and Market-Based Credit Models

In assessing credit risk, financial organizations usually consider two types of models: statistical models and market-based models [Taylor 2015]. Each of the methodologies offers different advantages for different aspects of credit analysis. On one hand, statistical models tend to use historical data to estimate future probabilities of default and so these models exploit numerous statistical techniques, including regression analysis, in order to analyze the interrelationships that exist among a set of different financial variables and the probability of default. They provide a structured way of credit risk analysis, and their application can be made to a wide range of borrowers, including small and medium-sized enterprises that do not have public debt pricing [Taylor 2015].

However, the limitation of any statistical model is rather obvious in its dependency upon historical data. The past may often prove to be not a very good predictor of the future, especially during periods of economic transition or turmoil [Taylor 2015]. Thus, statistical models might fall short when it comes to accurately conceiving the fluctuation in existing economic conditions or other borrower-specific factors leading to the possibility of default.

On the other hand, market-based models deduce data from real-time prices of financial instruments and provide dynamic views about credit risk. These models reflect the expectations of investors regarding the future performance of a firm and, hence, are susceptible to altering market conditions [Taylor 2015]. However, these models can only be applied to those companies with publicly traded financial instruments and secondly most models

assume that markets are efficient (that is, prevailing prices already reflect all information available). This, perhaps, may not be fully applicable under any and all circumstances, especially under conditions of market turbulence or irrational behavior on the part of investors.

In conclusion, both statistical and market-based models are complementary approaches toward credit risk assessment and although statistical models provide structured insights on the history of data, market-based models yield relevant information about current investor expectations and the dynamics of the market. These usually vary based on different circumstances at hand, that is, the nature of the assessed borrower, availability of data, and sophistication by the creditor.

Of all the problems that a credit risk management system would encounter, the challenge of catastrophic losses is the most critical, as written in [Brown and Moles 2014]. By their very nature, catastrophic losses are the ones that are highly infrequent but exceedingly large and thus often bring a financial institution to its knees.

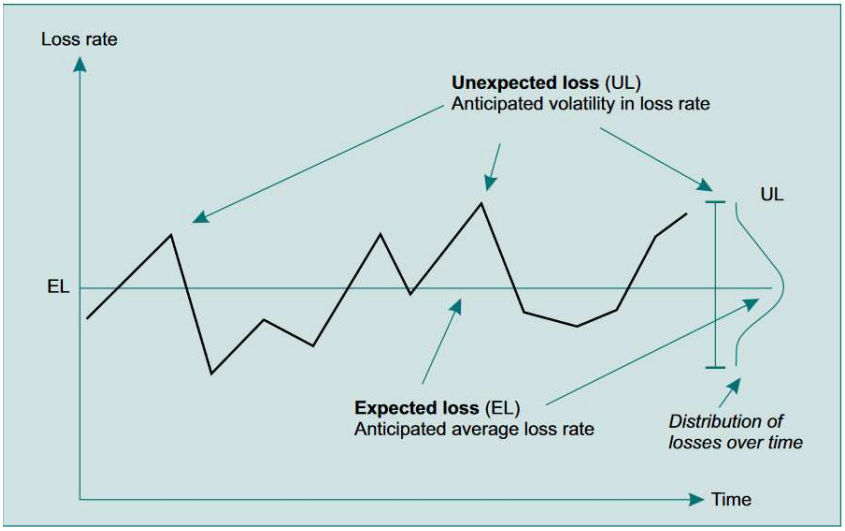


**Figure 1.3:** Distribution of Unexpected loss, from [Brown and Moles 2014]

These catastrophic losses can occur for a variety of reasons such as concentrated risk in specific sectors, geographic regions, or borrower categories,

when a large number of highly correlated credit events occur concurrently. For instance, a sharp economic downturn may trigger an across-industry wave of defaults that results in losses far beyond what is considered normal for a given risk.

In any case, the possibility of extreme losses places a premium on an appropriate and conservative credit risk management approach. It is not sufficient that financial institutions assess the expected loss and unexpected loss, the possibility of extreme events also needs to be considered and proactive mitigants put in place.

Controls that could be put in place to limit such potential huge losses include, according to [Brown and Moles 2014]:

- Portfolio Diversification: it implies limiting a concentration of risk by dispersing exposures across different industries or sectors, geographic regions, and types of borrowers. This would dampen sector-specific shocks or regional crises.

- Stress Testing: simulation under extreme market conditions, testing the robustness of the credit portfolio against large-sized shocks. Stress testing allows institutions to find their weak points and determine whether they are adequately prepared for such eventualities.

- Operational Continuity Planning: detailed planning for continuity in case of catastrophic losses. This would relate to protocols around communications, resource reallocation, and efforts at recovery to minimize disruption.

- Mechanisms of Risk Transfer: application of instruments like insurance or credit default swaps in order to pass on parts of risk to third-party agents. This provides added protection against the possibility of loss.

In summary, credit risk management is a hybrid process that integrates qualitative judgments and quantitative analysis in protecting the financial

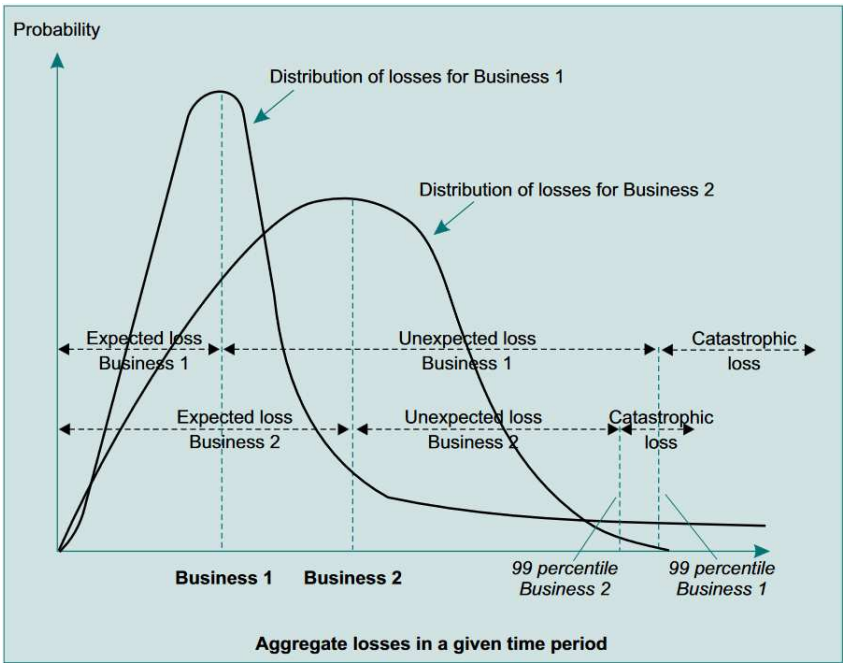


**Figure 1.4:** Example of loss distribution, from [Brown and Moles 2014]

institution from the loss inherent in borrower default. Expert-based approaches, credit scoring, and market-based techniques can all be applied to arrive at estimates of default and exposure to credit risk, thereby enhancing predictive capabilities.

## 1.4 Inflation

Inflation is the rate at which goods and services of an economy increase in price over a period of time. It is marked by a general increase in the price level, whereby purchasing power is reduced because money cannot buy as many goods or services as it did before because prices have risen. Preceding this phenomenon are various factors, such as increases in demand for goods and services, increases in the cost of production, or so-called "exogenous shocks" such as those seen in the increase in oil prices, which had such a great impact, especially in the 1970s [Perry et al. 1980]. During that decade, the American economy was gripped by a high rate of inflation brought partly by the oil price increases charged by OPEC.

Inflation is usually calculated by measuring the percentage change in the price level of a basket of goods and services over time. Here, we'll give some typical key formulas used to compute inflation, explaining their significance along the way.

- Consumer Price Index (CPI):

  The Consumer Price Index (CPI) is one of the most widely used measures of inflation. It tracks the cost of a fixed basket of goods and services typically purchased by households. The formula for calculating CPI is:

$$CPI = \frac{Costof BasketinCurrentYear}{Costof BasketinBaseYear} * 100$$

  Once the CPI is calculated, the inflation rate can be derived by comparing the CPI in two different periods [Perry et al. 1980]. The formula for the inflation rate (as a percentage) is:

$$InflationRate = \frac{CPI_{new} - CPI_{old}}{CPI_{old}} * 100$$

Where:

$CPI_{new}$ is the CPI for the current year or period.

$CPI_{old}$ is the CPI for the base year.

This formula gives us the percentage change in the price level, which is the rate of inflation.

- GDP deflator:

  Another way to measure inflation is through the GDP deflator, which reflects the price level of all domestically produced goods and services [Perry et al. 1980]. Unlike CPI, the GDP deflator is not based on a fixed basket but reflects the prices of all final goods and services. The formula is:

$$GDPDeflator = \frac{NominalGDP}{RealGDP} * 100$$

  Where:

  $NominalGDP$ is the total value of goods and services produced, measured at current prices.

  $RealGDP$ is the total value of goods and services produced, adjusted for inflation (measured in constant prices).

  Inflation can also be calculated using the GDP deflator in two different periods, using the formula:

$$InflationRate = \frac{GDPDeflator_{new} - GDPDeflator_{old}}{GDPDeflator_{old}} * 100$$

  This method accounts for price changes in the entire economy rather than just consumer goods.

One of the central features of inflation to emerge from the analysis is the concept of "inertia", as reported in [Perry et al. 1980]. This concept refers

to the tendency of inflation, once it takes hold, to persist over time even when economic conditions change. A good example of this inertia can be explained by the fact that workers and companies, while negotiating wages, take into account not only current inflation but also past inflation and their future expectations. This is expressed through the "wage norm" [Perry et al. 1980], a concept by economist Perry that means the rate at which wage growth is expected by workers and businesses based on past inflation and union contracts. When these expectations are built in, inflation becomes hard to contain since workers, even during periods of economic recession, are very hesitant to see their nominal wages diminished [Packer and Zhu 2012].

There are several economic policies available to fight inflation and one of these is maintaining the unemployment rate at a high level. The rationale behind it is that with high unemployment, the demand for labor becomes low, hence minimizing the pressure to increase wages and, consequently, prices. However, this approach entails high costs of production and employment. Another instrument of policy is that of restrictive fiscal and monetary policy, whereby public spending is reduced or interest rates are increased. It may try to cool down the economy and dampen inflationary pressures but will be effective only to the extent that it can actually alter expectations on price growth [Perry et al. 1980].

Some economists put forward the view that, if an announcement of restrictive policy is made in a credible and consistent way, then it could change expectations about inflation by economic agents and result in quicker attenuation. But the secret to this tactic is credibility: unless firms and employees believe that the government or central bank will keep this stance, the policy will have little effect. Yet another possible avenue is through the use of so-called income policies, such as wage and price controls or tax incentives to temper their rise [Packer and Zhu 2012]. The goal of these instruments is to

attempt directly to affect the processes of wage and price setting, but how effective and pervasive that influence has been is, however, open to debate.

With the different theoretical approaches to studying inflation, there would be two that stand out: the Keynesian approach and the monetarist/neoclassical approach [Perry et al. 1980].

The Keynesian approach, derived from theories by John Maynard Keynes, is said to view inflation as being determined both from demand and supply. In low unemployment conditions, the demand for goods and services increases, wages increase, and companies pass higher labor costs by increasing prices to feed inflation. Supply-side external shocks involve commodity price increases, such as oil, which can also create a general price increase. The Keynesian approach describes a relationship between inflation and unemployment using the Phillips curve, where increased employment, in the short run, tends to drive up wages and hence prices [Perry et al. 1980]. On the other hand, this approach recognizes that the long-term characteristics of the curve are unstable since inflationary inertia provides an obstacle in reversing the trending increase in inflation once it has taken hold.

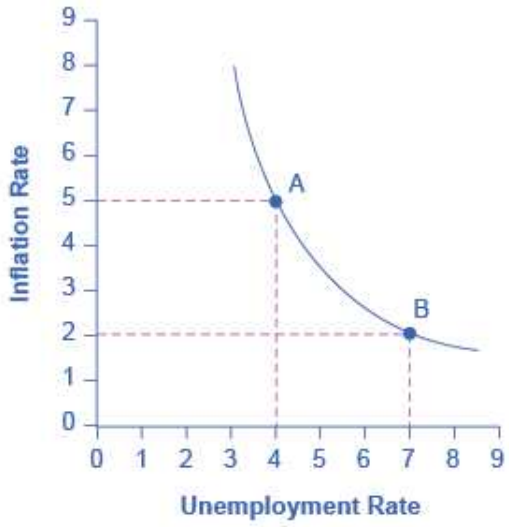


**Figure 1.5:** Relation between Inflation and Unemployment , from [Perry et al. 1980]

On the other hand, the monetarist and neoclassical approach sets great store by the part played by the money supply. In this view, the root cause of inflation is monetary: inflating prices are the result of too rapid an increase in the amount of money in circulation. Monetarists, with their theory inspired by Milton Friedman, call for price stability only if the money supply increases constantly and at a moderate rate. Neoclassical models are thus then based on rational expectations, in which the people form their future expectations rationally and with all the available information. These models suppose that inflation cannot be persistent, since economic agents do not expect further price increases and the markets have a natural tendency towards full employment equilibrium.

It has also been put that the monetarist approach is not in a position to explain such a complex economic occurrence as the oil shocks in the 1970s [Perry et al. 1980]. During that period, there was a rapid rise in inflation, related to the explosive growth in the price of oil, while money supply growth was too small in order to explain the magnitude of the inflationary phenomenon. The rational expectations hypothesis has also been challenged, because, in the real world, people are typically denied the relevant information to make perfect predictions, and systematic errors are pervasive.

The numerous theoretical perspectives on inflation determine the various causes it may take and ways of its potential overcoming. The Keynesian approach views aggregate demand and wage rigidity as the main causes of this malady. Instead monetarists and neoclassicals stress money and rational expectations [Perry et al. 1980].

## 1.5 Unemployment

Unemployment is some sort of economic phenomenon that manifests when people of working age are willing to work but cannot find any work. It is a state of disequilibrium in the labor market (a mismatch in the supply

and demand for human resources, which, though available, are not utilized productively). This may come in several forms, but the most objectionable is involuntary unemployment since it clearly shows that while people want to work, they just cannot because there is not enough labor demand [Packer and Zhu 2012].

This macroeconomic phenomenon is correlated with credit defaults, as situations with high unemployment or job insecurity increase uncertainty, which negatively impacts loan repayment. In such conditions, borrowers may struggle to meet their obligations, leading to a rise in loan defaults ["Global financial development report 2019/2020: Bank regulation and supervision a decade after the global financial crisis" 2019].

Consequences of unemployment on the economy and the social sphere are extremely serious. Economically, it means a productivity loss due to potentially useful resources being idle. This factor not only reflects in a reduction of overall output but also contributes to an increase in social costs and reduced general well-being among the population. In addition, involuntary unemployment is often accompanied by deterioration in motivation and skills among the workers themselves, which further creates a self-perpetuating vicious circle that makes their return to the labor market much more difficult.

Apart from that, unemployment strongly influences social life: it leads to the growth of impoverished families, increases income disparity, worsens the living standards of the population, and even causes the deterioration of mental health. The lack of a job means unemployed people will lose their principal source of subsistence, which could bring about a decline in the living conditions and mental health of one ["Global financial development report 2019/2020: Bank regulation and supervision a decade after the global financial crisis" 2019]. These family consequences just make for an environment of tension and uncertainty that could badly affect future generations.

The causes of unemployment can be various and complex and include

economic shocks, like crises or recessions, which suddenly lower the demand for labor, labor rigidities, engendered by restrictive regulations, curtail hiring and firing options available to firms and make adjustments in response to changing economic conditions hard. Apart from that, structural unemployment may also be contributed by a mismatch in the qualifications relevant to the company's needs [Packer and Zhu 2012].

Long-term unemployment is a very disturbing phenomenon. People who stay out of work for a long period usually lose their self-confidence and may give up on any active job searching, therefore worsening the situation even more. The excessive tax burden can contribute to discouraging hiring, pushing companies towards other alternatives such as tax evasion or hidden hiring, distorting the labor market.

Basically, unemployment can be seen as an economic issue with strong social overtones. For its analysis, it requires a multi-factor approach that could consider the amount of reasons and results it implies. The importance of doing this is to find a way of having good policies that could help in overcoming this phenomenon in order to boost employment and increase the living conditions for the population ["Global financial development report 2019/2020: Bank regulation and supervision a decade after the global financial crisis" 2019].

Unemployment in Italy represents a complex phenomenon caused by so many different factors that one is hard put to single out the priority cause.

One of the structural factors is the dualism of the labor market, considering permanent workers, in some respects protected by various safeguards, and precarious workers, without effective protections. This dualism has been worsened by reforms in favor of flexibility, to which young people are extremely vulnerable, especially during an economic crisis. Companies avoid transforming fixed-term contracts into permanent ones, with results such as inefficient turnover and little or no opportunities for training for temporary

workers [Packer and Zhu 2012].

Another critical factor is the incomplete supply of opportunities to receive training. The weak development of education and vocational courses, especially in training specifically for young individuals, closes them off to labor market opportunities. On the other hand, firms, when seeking recruits, favor those with previous experience above all else; this forms a sort of vicious circle, whereby the youngest find it really hard to get their first job ["Global financial development report 2019/2020: Bank regulation and supervision a decade after the global financial crisis" 2019]. Besides, there are great territorial disparities: southern Italy shows unemployment rates far superior to those of the North and the Center. The cause of such a gap has to do with the far weaker economic structure, the low degree of industrialization, and the underdeveloped entrepreneurial fabric. Still, there are some important gender disparities: Italy has one of the lowest female employment rates in Europe because of cultural and structural barriers to women's participation in the labor market.

Other important cyclical factors also contribute to unemployment. A financial crisis, such as the one in 2008, and then the sovereign debt crisis in Europe, hit Italian employment with particular violence. Contraction in both internal and external demand has brought about a reduction in production levels and increased layoffs. The flexibility of the labor market helps economic growth when the expansion is going on but turns the employment fragile during recessions.

Cultural factors also exert their influence on unemployment. For example, the low geographic mobility of Italian workers is culturally and familistically pushed toward residing in places that decrease their mobility to other, more favorable labor markets. The cultural resistance to change and difficulty in adapting to new labor market demands may also further contribute to unemployment.

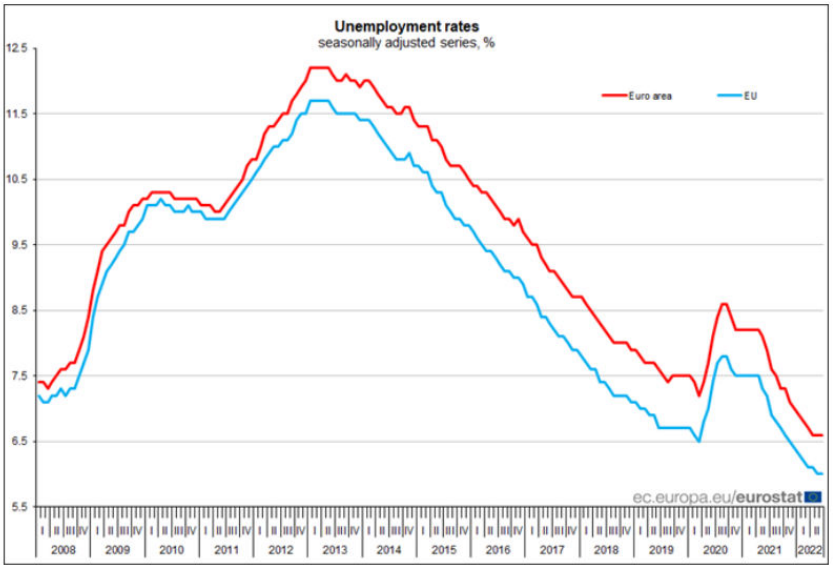


**Figure 1.6:** Unemployment rate from 2008 to 2022, from [Perry et al. 1980]

Last but not least, technological unemployment is another more significant reason. The technological advancement with automation and artificial intelligence may replace workers with machines. However, that also opens possibilities for new opportunities in emerging sectors that require innovative skills [Stulz 2001].

In a nutshell, Italian unemployment is a multilayered problem with strong structural, cyclical, and cultural roots. For such a challenge to be tackled, a multidimensional approach has to include labor market reforms to reduce dualism, investments in appropriate education and training, active policies for the unemployed, measures of support for enterprise economic growth, incentives to geographic mobility, and innovations that help in shifting to new sectors.

## 1.6 House prices

House prices have a strong correlation with the capacity of the borrowers to repay their debts. As shown in [Tajik et al. 2015] ,research conducted in

the USA, there are relationships between prices and NLP levels and this can affect the probability of default of a counterpart, so it can be an important feature to take into account. The paper of [Tajik et al. 2015] provides a clear analysis of how house price fluctuations affect the credit risk of banks in the United States, focusing on NPLs, since both are deemed critical indicators of stability for the banking sector and their loan quality.

The study has shown that house price dynamics shape not only the credit risk landscape but also have a cyclical, asymmetric relationship with increased or decreased house prices and changes in NPL levels [Tajik et al. 2015]. House price booms positively influence borrowers' debt-repaying capability as the value of collaterals, real estate, increases. In return, this minimizes the possibility of defaults by many debtors. Therefore, a housing market boom translates to a low probability of default rates for banks, meaning reduced nonperforming loans or bad debts.

On the contrary, a fall in house prices reduces the collaterals' value drastically and therefore paves the way for more defaults due to lower equity in properties on the part of the borrowers. This decrease in the value of collateral implies that, during this period, banks have an exceptionally high rate of default, which significantly heightens the volume of NPLs and increases the severity of credit risk.

Some of the results are in Figure 1.7.

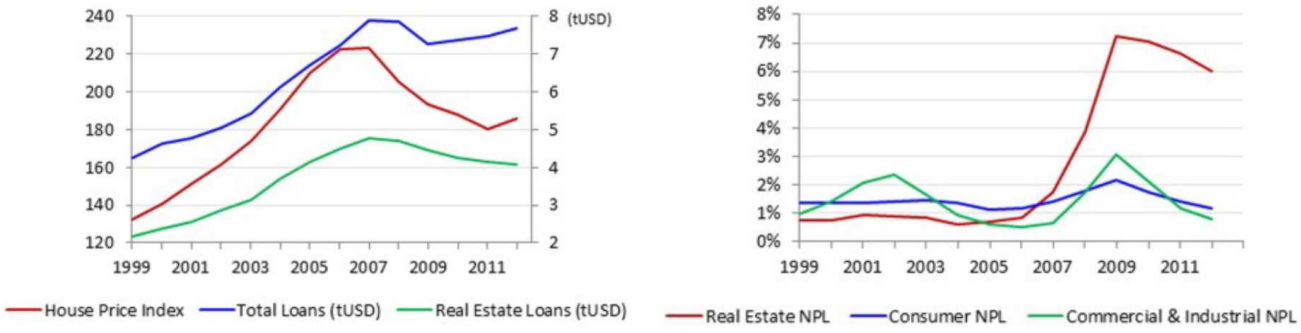


**Figure 1.7:** House prices and NLP, from [Tajik et al. 2015]

These findings support the fact that the relationship between house prices and credit risk is strongly asymmetric: while house price booms give only a moderate decrease in credit risk, the fallen house price impact disproportionately hits hard Tajik et al. 2015. That asymmetry is greatest in the case of a downturn: borrowers are less able to use other assets or increased income as offsets against the loss in property value and therefore experience a significant rise in the number of defaults, as well as an increase in NPL throughout the banking institutions. This is a useful insight for it allows the understanding that in respect to real estate loans, credit risk can build up during booms in housing markets and then realize as sudden large losses in banks when these markets turn downwards.

To better understand this dynamic, the study of [Tajik et al. 2015] further decomposes the relationship between house prices and credit risk by loan type and bank type. Real estate loans are particularly exposed to a decline in house prices since they are dominated by real estate collateral. As house prices fall, the value of the collateral that supports these loans deteriorates and thus precipitates defaults from borrowers in financial distress. Commercial and consumer loans, by contrast, are typically much less sensitive to a house price downturn since they are often unsecured or held against non-real-estate assets, which by their very nature would render them more resistant to the housing market. This further suggests that banks should closely monitor their real estate loan portfolios, as high concentrations in those types of loans could increase their vulnerability to financial stress in the event of a housing market downturn [Packer and Zhu 2012].

The comparison of the effect that house price changes have on different types of banking institutions was made further for Commercial Banks (CB) and Savings Institutions (SI). In as much as SIs pay a great deal of attention to mortgage lending, alongside home ownership, it would appear that CBs are actually more sensitive to the decline in house prices [Ongena and Smith

2001]. This heightened sensitivity among CBs is likely because their scope of lending is so much wider, and hence includes more riskier types of real estate loans than just residential. As a result, when house prices fall, CBs tend to suffer more from rising NPLs, possibly because they do not have the same degree of specialization in residential mortgage risk management as SIs do [Tajik et al. 2015]. This suggests that institutional differences in loan portfolio composition and risk management practices are the important determinants of a bank's vulnerability to changes in house prices.

The results of the present study have some significant implications for financial regulation and macroprudential policy. House prices are thus a significant macroprudential indicator, with the established transmission between house prices and credit risk (a function that carries information about developments in future credit risk). On the other hand, against the backdrop of falling house prices, regulatory agencies may be justified in using countercyclical capital buffers, particularly for those banks with high concentrations of real estate loans, to give them a financial cushion against such loan losses. This study [Tajik et al. 2015], therefore, puts additional weight on active monitoring by banks and regulators of the composition of loan portfolios, particularly exposure to the real estate sector, and also being prepared for increasing credit risk associated with economic downturns. It, therefore, creates a clear route through which financial institutions and policymakers would be in a position to apply different ways of enhancing financial stability in order to protect the banking industry from extreme disruptions that are likely to come with changes in the housing market, basing their approach on the understanding that there is cyclicality and asymmetry in the relationship between house prices and credit risk [Packer and Zhu 2012].

## 1.7 Interest rate curve

The paper from [González-Aguado and Suarez 2015] investigates the relationship between interest rates, political incentives, and corporate credit risk. Most political decisions involving interest rates are based on stimulating economic growth and stabilizing it: the process can only create an enabling environment that has low borrowing costs, hence encouraging firms to increase their leverage as a way of influencing credit risk over time.

Some of the key drivers of this interrelationship include the "quest for yield." Low interest rates, which are sometimes politically driven to stimulate economic activity, make firms and financial institutions seek higher returns by buying riskier assets [González-Aguado and Suarez 2015]. After the 2008 financial crisis, for example, interest rates were kept extremely low to ensure that borrowers could borrow more and invest more. While this policy supported short-term growth, it also precipitated higher leverage as firms sought to capitalize on the low-cost borrowing environment. Increased leverage heightened the vulnerability of the financial system to firms and financial institutions taking riskier investments in their search for returns above the low risk-free rate.

In the paper of [González-Aguado and Suarez 2015] we can also find how firms respond to interest rate changes according to factors such as firm age, leverage level, and state of the economy. Firms are typically born highly leveraged because of the insufficiency of their internal funds and are forced to grow through debt. As they grow older, they gradually deleverage through retaining earnings towards a target that would minimize the risk of default.

When interest rates are cut, mature firms rapidly move to a new, higher target leverage and increase debt to take advantage of the lower cost of borrowing. Younger firms also gain from lower rates because they can reduce burdens from debts accumulated earlier with greater ease and thereby achieve greater financial stability [Al-Gunaid et al. 2021].

On the other hand, in the case of a rate increase, highly leveraged firms, usually younger ones, suffer due to higher costs of debt servicing. The outcome is increased short-term risk of default because those firms are not in a position to deal with the burden caused by the rate increase [González-Aguado and Suarez 2015].

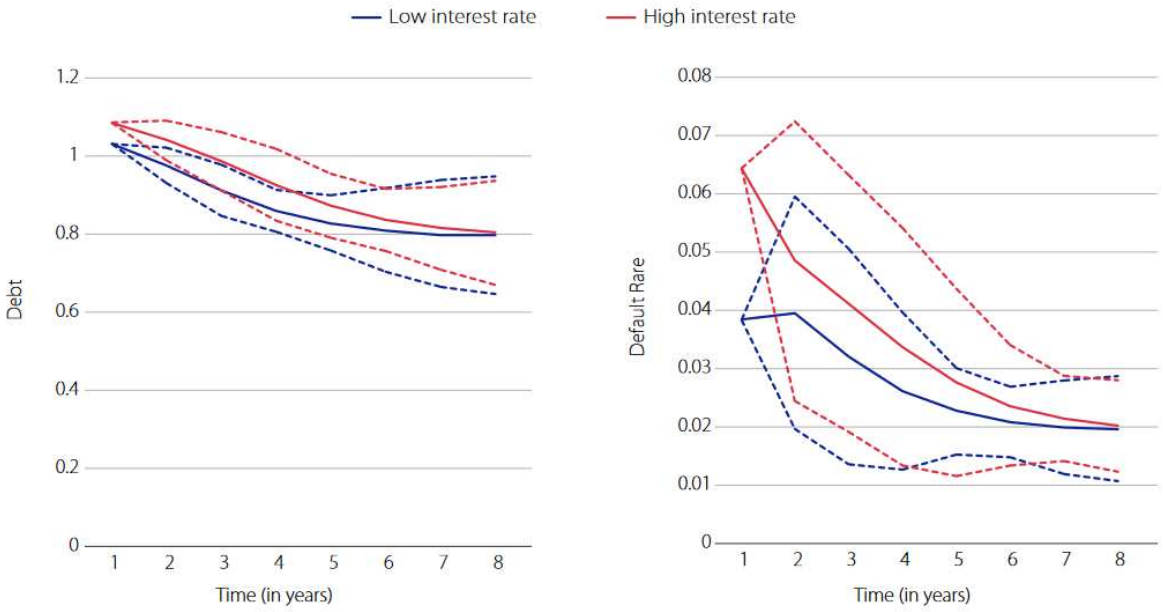


**Figure 1.8:** Default related to Interest rate shifts

Rate increases also put higher stress on highly leveraged companies, whose credit risk increases immediately while their mature counterparts enjoy greater flexibility in reducing target leverage. Low-interest policies that are politically motivated will give temporary stimulus but create aggregate credit risk over time. Rate reductions increase the target leverage of firms, driving up the default rate as leverage increases throughout the market. On the other hand, while rate rises can raise temporary credit risk for the highly leveraged firms, they tend to give lower overall credit risks as the firms are urged to reduce leverage. It is this relationship between interest rates and credit risk that has made interest rate policy significant from a macroprudential perspective, as reported in [González-Aguado and Suarez 2015]. Policies

keeping low rates for longer create vulnerabilities in the financial system due to excessive risk-taking.

While this may look attractive to policymakers for generating growth, there are long-term risks associated with this type of strategy that need to be weighed. Because the model says that an optimal interest rate policy needs balance, an extremely low interest rate may foster unsustainable leverage and credit risk across the economy. This dynamic reinforces the need for judicious consideration of the "risk-taking channel" of monetary policy, whereby low rates stimulate growth and may concurrently drive increased credit risk.

# Chapter 2

# Machine learning

Machine learning can be seen as one of the arms through which computers are able to learn and get better at performing specific tasks without explicit programming. The general idea behind machine learning involves the use of algorithms that identify patterns within large datasets so that the system can eventually make decisions and predictions independently [Khandani, A. Kim, and Lo 2010]. It is dependent on the training of mathematical models with historical data, finding applications in speech recognition, driving autonomously, medical diagnostics, and marketing. It also finds extensive applications amongst financial institutions and banks in the form of market research on financial market forecasting and credit risk assessment, thus enabling better decisions on risk management and investment decisions [Matz and Luo 2021]. Here we want to present two wide-used models which we want to use for the pourpose of this thesis.

## 2.1 Decision trees

Decision trees are among the most interpretable and ubiquitous models used for classification and predictive modeling, as in [Rokach and Maimon 2005]. These are applicable in machine learning, statistics, and data mining. In

principle, decision trees operate by recursively partitioning the example space according to specific attribute tests in such a way that during classification, instances run through a tree-like structure that governs their class membership.

The structure of a Decision Tree may be defined as the set of nodes organized in an ordered tree-like structure from the root, normally known as the first test. Each internal node denotes a test on some attribute, while each leaf is labeled with a class or final decision [Khandani, A. Kim, and Lo 2010]. En-route from top to bottom, the instance space is divided recursively into smaller subsets until the leaves are reached, indicating the classification of the instance.

Decision Tree Building Process Decision trees are constructed by using some algorithms that do the splitting in a recursive way. Some of the most well-known algorithms include ID3, C4.5, and CART [Rokach and Maimon 2005]; all of them use the "top-down" approach: they start with the splitting of the whole data set and keep on dividing it until some predefined stopping criteria are reached. The basis for splitting data, in most cases, would normally be an impurity measure, which gives an idea of the homogeneity of the data once a split has been applied [Khandani, A. Kim, and Lo 2010]. The best attribute to choose will be the one that most minimizes the uncertainty of the data.

One important aspect in decision tree building involves the determination of the best split for every node. A number of measures can be used to assess the quality of a split:

- Information Gain: Computed with respect to entropy, it is the reduction in impurity within a split;

- Gini Index: A measure of impurity. It describes how well a split is able to separate classes. This is usually used in binary trees;

- Gain Ratio: A normalized information gain. This was devised so that an attribute with a great deal of unique values would not be favored overtly.

Moreover here are some examples of the implementation of Decision trees, as written in [Rokach and Maimon 2005]:

- ID3: A simple algorithm that carries out a split of the data based on information gain. It stops when all the instances in the subset belong to the same class, or when further splits will not improve the classificationè;

- C4.5: The evolution of ID3, which makes use of the gain ratio in order to enhance its predecessor's splits. It also handles numeric attributes and missing values much better;

- CART: It is a renowned algorithm that constructs classification and regression trees alike by using the Gini index with a complexity-based pruning strategy.

### 2.1.1 Prunear Decision Trees

One of the big problems while building decision trees is overfitting. When the size of a tree becomes inordinately large, it fits too closely to the training data, and its generalization for new, unseen data is poor. To overcome this problem, pruning methods are used: it simplifies the tree by removing sections which do not play a crucial role in the performance of the overall tree model.

One popular method is called cost-complexity pruning, a method for finding an optimal balance between the size of the tree and some estimate of generalization error. The procedure examines a tree, and for any given branch, determines whether removing it can result in a model that better

generalizes outside the training data, even at some expense of the performance on the training data itself.

The alternative is reduced error pruning, in which nodes are only removed if such removal does not affect the generalization capability of the tree on an independent validation set. That way, decisions take into account not only the performance of the training set but also that of unseen data-the key to the better generalization of the tree.

Finally, pessimistic pruning uses an internal estimate of the error rate (called pessimistic error estimate) without holding out a separate validation set. The method assumes the true error rate is actually higher than that on the training data and prunes the tree thus. This method simplifies the model by accounting for uncertainty in the error estimate; the goal is to avoid overfitting without extra data on validation.

### 2.1.2 Advantages and Disadvantages of Decision Trees

As reported in [Rokach and Maimon 2005], decision trees have a host of advantages working toward making them quite popular. First, the output of decision trees is relatively simple to comprehend and infer from. This is because decision trees are, to a large extent, non-expert-friendly, especially if they are not too deep. Besides, they can handle numeric as well as categorical attributes.

Another key characteristic of decision trees is that they are non-parametric: this means no assumptions of the underlying distribution of the data are made, which allows flexibility in their application to a wide range of problems and domains. It is not required to adapt the model to a certain data type.

Decision trees, however, have also some drawbacks. A common problem with these is that they can be so vulnerable to overfitting (as written in [Rokach and Maimon 2005]), especially in cases when the tree has become

very large and pruning does not go well. Overfitting occurs when the model fits too closely to the training data, to the detriment of generalizing to new data.

Decision trees do quite well where highly relevant attributes are available, but they fail to model more complex interactions between multiple variables. Apart from that, they tend to be sensitive to noisy data and irrelevant attributes of datasets, by which their performance is affected.

Some of these deficiencies were overcome by variants of the standard decision trees. A very well-known variant, for example, is the oblivious decision tree, in which all nodes at one level of an oblivious tree test the same attribute. This makes the interpretation easier-the structure of this type of tree becomes simpler to trace and manipulate. A different extension concerns the so-called fuzzy decision trees where an instance can be assigned to more than one branch at the same time, each with a different degree of confidence. This, in turn, helps to deal with uncertainty in data more effectively by flexible treatment of ambiguous or overlapping information.

Decision trees keep evolving with the growth of the volume and complexity of a classification problem. Recent algorithmic developments deal with methods for handling larger datasets and a way to incrementally update a decision tree as new data comes in without having to rebuild the entire tree from scratch. These enhancements allow for scalability and adaptation in decision trees according to today's demands.

## 2.2 Random Forest

Random forests are a type of ensemble learning where multiple decision trees are used together to give a more accurate result with less overfitting. In general, they are really useful at regression and classification tasks [Matz and Luo 2021]. Mathematically, they aggregate the predictions over many trees such that each tree is grown from a bootstrapped sample of the training

data and a random subset of the features. This technique will make a model robust, shrinking the high variance associated with single decision trees,as written in [Biau and Scornet 2016].

Assume a generic supervised learning problem with given training set $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $X_i \in \mathbb{R}^p$ is the feature vector of $i$-th sample, and $Y_i \in \mathbb{R}$ for regression problem, or $Y_i \in \{1, 2, \ldots, K\}$ for classification problem is the target label.

The objective is to estimate a function ( f(X) ) that minimizes some error metric (e.g., mean squared error for regression or classification error for classification tasks). In the case of the random forest approach, this function is expressed as the average of many decision trees. For regression, the random forest predictor ($\hat{f}(x)$) takes the form:

$$\hat{f}(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(x), \tag{2.1}$$

where ($T_m(x)$) is the prediction of the (m)-th tree, and (M) is the total number of trees. For classification, the output is the class that receives the majority of votes from all trees:

$$\hat{f}(x) = \text{argmax}_k \left( \frac{1}{M} \sum_{m=1}^{M} 1(T_m(x) = k) \right), \tag{2.2}$$

where ($1(\cdot)$) is an indicator function which returns 1 if ($T_m(x) = k$), and 0 otherwise and $k$ represents one of the possible classes.

Each tree ($T_m(x)$) is grown based on a bootstrapped sample ($D_m$) of the training data [Biau and Scornet 2016]. That is, to grow each tree, a random sample of $n$ observations with replacement is drawn from the training set. Approximately $\frac{2}{3}$ of the original data points will appear in each bootstrapped sample. The remaining $\frac{1}{3}$ form the test set, which can be used to estimate prediction error.

Recursive partitioning of the feature space is done in building the tree. At each node of the tree, a random subset of size $m_{\text{try}}$ of the total $p$ features

are chosen. From that subset, the best split - according to some criterion - such as Gini impurity for classification, or MSE for regression is used.

In classification problems, for a set of data points $\{(X_i, Y_i)\}$ at a node, Gini impurity $I_G$ is defined as,

$$I_G = 1 - \sum_{k=1}^{K} p_k^2 \tag{2.3}$$

where $p_k$ is the fraction of data points at the node that belongs to class $k$. On the other hand, the best split in a classification problem will be chosen by returning the split which minimizes a weighted Gini impurity across the two child nodes.

The best split in regression aims at minimizing the MSE, defined as:

$$I_{\text{MSE}} = \frac{1}{n_{\text{left}}} \sum_{i \in \text{left}} (Y_i - \hat{Y}_{\text{left}})^2 + \frac{1}{n_{\text{right}}} \sum_{i \in \text{right}} (Y_i - \hat{Y}_{\text{right}})^2, \tag{2.4}$$

where $\hat{Y}_{\text{left}}$ and $\hat{Y}_{\text{right}}$ are the average target values of the left and right child nodes after the split and $n_{\text{left}}$ and $n_{\text{right}}$ are the number of samples in each child node.

Once all trees are constructed, the random forest combines their predictions. For regression, the predicted value at $x$ is the average of the outputs of all the trees:

$$\hat{f}(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(x). \tag{2.5}$$

For classification, each tree casts a "vote" for a class, and the final prediction is the class with the most votes:

$$\hat{f}(x) = \text{mode } \{T_1(x), T_2(x), \ldots, T_M(x)\}. \tag{2.6}$$

### 2.2.1 Parameter Tuning

In [Biau and Scornet 2016] there iare also some major influential parameters in the performance of a random forest, which have to be judiciously tuned for optimum performance, include

- Number of trees ($M$): With an increase in the number of trees, the variance decreases and hence the prediction becomes more stable. From the law of large numbers, as $M \to \infty$, the variance of the random forest diminishes and the predictions converge to a stable value:

  It follows that:

$$\lim_{M \to \infty} \hat{f}(x) = E[T(x)]. \tag{2.7}$$

  For most practical purposes, $M$ is selected to balance computational resources against performance. Typical values range between 100 and 1000.

- Number of features per split ($m\ try$): This is a factor that introduces randomness in the feature selection at each node. By default, for classification $m\ try = p$ and for regression $m\ try = p^3$. The smaller the value $m\ try$, the more diverse the trees can get, hence the less correlation among trees, so improvement in generalization.

- Minimum samples per leaf (min_samples_leaf): This makes a preventive measure against trees growing too deep, hence causing overfitting. Moving to a larger value for min_samples_leaf will force a tree to generalize more, which will decrease the variance with the potential disadvantage of having higher bias.

- Maximum depth (max_depth): Some algorithms also make use of a stopping criterion that constrains the depth of the trees. Shallow trees

capture less complexity in the data; thus, the probability to memorize noise is reduced.

### 2.2.2 Test Error and Variable Importance

Random forests naturally estimate performance via the test error or out-of-bag error [Biau and Scornet 2016]. Due to the process of bootstrapping, about $\frac{1}{3}$ of the training data are not used to build each tree. One could then test the performance of the tree based on the left-out OOB data and the error is taken as the average error across all trees with the OOB data points.

In addition, random forests allow for the calculation of variable importance, which quantifies the contribution of each feature towards the performance of the model. It does this by randomly permuting the values of a feature in the OOB data and then calculating the resulting increase in error in prediction. The more error increases when a certain feature is permuted, the more important it is considered to be [Matz and Luo 2021].

There are also some techniques used in machine learning in order to obtain better results. These include metodologies such as cross-validation in order to understand the impact of different combinations of parameters [Nitesh 2002]. In practice, hyperparameter tuning often focuses on the number of trees, $m_{\text{try}}$, and depth or minimum samples per leaf as a means to achieve an ideal bias-variance trade-off [Matz and Luo 2021]. The out-of-bag error furnishes an inbuilt mechanism of model performance evaluation during its training and makes certain fine tunings easier without the need for a separate validation set.

Random forests are among the most flexible, robust, and accurate machine learning models developed based on the aggregation of many decision trees. The ability of this model to handle high-dimensional data, estimate the importance of variables, and avoid overfitting using randomness in both data and feature selection is what makes it very powerful.

# Chapter 3

# Machine learning in Credit Risk

## 3.1 Machine learning for Credit risk

In evaluating credit risks, financial institutions have generally followed the model of the "Five Cs of Credit", as reported in [Bazarbash 2019]. These five elements provide a structured framework for investigating the capability of a borrower to repay his loan and, accordingly, the risk involved in extending the credit. The Five Cs of Credit are:

- **Capacity**

  The term capacity refers to the borrower's financial ability to repay the loan; their income, expenses, and current commitments in respect to debt are considered in order to assess that ability. The debt-to-income ratio is one of the major precursors that would show this capacity. A fairly low DTI, especially when the economy is stable, would reveal that a borrower is likely to remain solvent even if the economic situation deteriorates [Thomas, Crook, and Edelman 2017]. This encompasses a wider range of information, which includes non-traditional factors that

are daily spending and purchasing behaviors, through the utilization of machine learning approaches and hence yielding a more accurate look into the financial situation of the borrower.

- **Capital Structure**

  Capital structure is the composition of the company's liabilities. A higher capital ratio, or equity that makes up a greater part than debt, is normally a good thing, for it suggests greater financial stability and incentivizes management to make a profit more. Through the machine learning analysis, one can examine specific financial metrics, which include cash flows and profitability in order to arrive at more granular predictions regarding business capability in paying obligations.

- **Collateral**

  Collateral is related to the value of the assets that the borrower puts up as security for the loan. The greater the collateral, the greater is the security of lenders. Through ML analysis, the value of the collateral can be estimated based on market factors, economic trends, and historical data related to similar assets, thus increasing the chances of loan recovery if the borrower defaults.

- **Character**

  Character: The credit history of the borrower, such as late payments, defaults, and legal issues. A high character score would indicate the borrower is deemed reliable and responsible. In fact, one could use machine learning algorithms to look at alternative data, including but not limited to payment history and even social media activity, to infer a more holistic view of the behavior and creditworthiness of the borrower.

- **Conditions**

  Conditions can be defined as those external factors that may affect

the borrower's ability to repay. It includes the state of the economy, the state of the industry, and geographical context. Machine learning models will parse extensive volumes of data on economic indicators to determine when emerging risks or opportunities may affect the borrower's ability to pay.

### 3.1.1 Machine Learning Methods Used in Credit

The application of machine learning methods in credit risk assessment has several advantages over traditional statistical models and brings significant improvement in the ability to assess and manage credit risk [Thomas, Crook, and Edelman 2017]. Following are some of the commonly used machine learning techniques, what each does, and its implications in credit risk assessment [Bazarbash 2019].

- **Decision Trees**

  Decision trees, also known as decision diagrams or trees, are tree-like model structures in which decisions and their possible consequences, including chance event outcomes, resource costs, and utility of outcomes, are depicted in a tree-like diagram [Rokach and Maimon 2005]. They can be applied to both classification tasks, such as the probability of a borrower's default, and regression tasks, such as LGD amount prediction.

  In credit risk assessment, decision trees are able to make predictions based on variables including Debt-to-Income and Loan-to-Value ratios. As an example, a decision tree might segment borrowers into classes of risk based on their financial attributes and allow lenders to make informed decisions [Bazarbash 2019].

  The main advantage of decision trees is interpretability, that is the fact that stakeholders can actually see how decisions are made adds to the

transparency of the decision-making process. However, decision trees are complex and tend to overfit the training data, especially for large sets of data points. This overfitting then plays a major role in the poor performance of the model upon application on unseen and new data [Frolov and Lavrentyeva 2019].

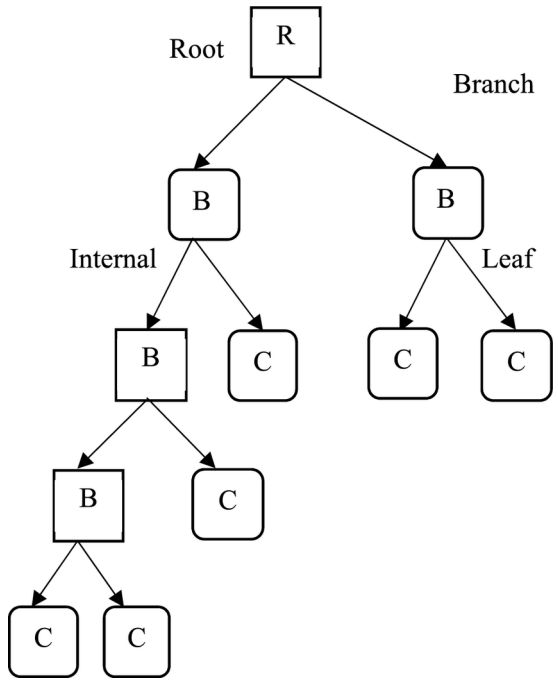


**Figure 3.1:** Decision tree example

- **Random Forest**

  Random Forest is an ensemble learning technique that generates a multitude of decision trees by taking bootstrapped samples of the data [Thomas, Crook, and Edelman 2017]. Each of these trees gets trained on a random subset of the features, and the final predictions are made through aggregation (most votes or averaging) over all trees.

  In credit risk assessment, this method of Random Forest can be used to enhance the robustness of predictions that assist lenders in ascertaining the risk associated with borrowers, as cited in [Bazarbash 2019].

Also, feature importance gives insight into which variables influence the model the most to predict the credit risk. Bagging or bootstrap aggregating helps to reduce the variance of the model and generally avoids overfitting issues, hence making it more generalizable on newer data.

Random Forests can be useful when there is a large dataset with many features, as they retain accuracy without falling into some of the pitfalls of single decision trees [Biau and Scornet 2016]. On the other hand, while Random Forest improves performance on a prediction task, it is generally less interpretable than a single decision tree because of its aggregated nature, obscuring the specific reasons behind individual predictions.

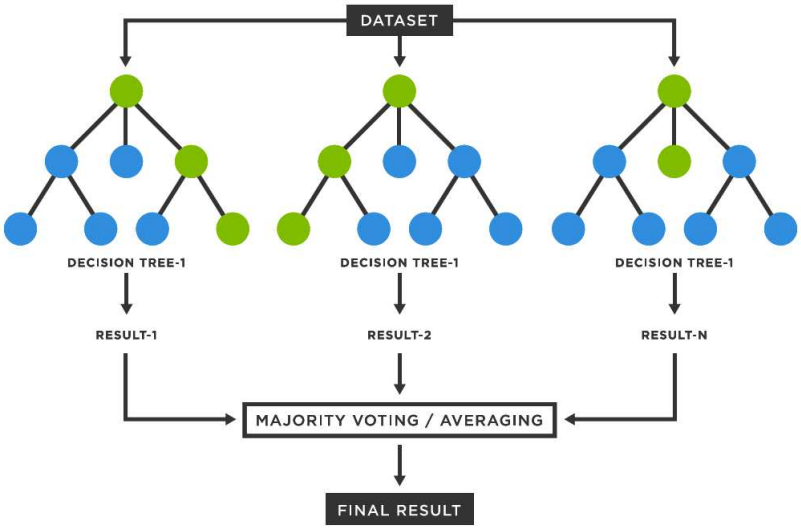


**Figure 3.2:** Random Forest implementation

- **Gradient Boosting Decision Tree (GBDT)**

  GBDT generates trees in sequence, with each new tree built to correct errors in the previous ones. Its iterative nature allows for natural gradual improvement of the model's accuracy. GBDT is widely recognized

for its predictive power and finds extensive application in credit rating and risk assessment models [Bazarbash 2019].

It can effectively handle various types of data and provide insight into the borrower's risk profile. Among the key strengths of GBDT are its high accuracy [Frolov and Lavrentyeva 2019]. Since each new tree is focused on the correction of errors of trees that came before, performance is optimized and hence effective in complex datasets. However, GBDT models are usually computationally expensive and sensitive to hyperparameters, which have to be tuned cautiously in order to avoid overfitting. Besides, though GBDT models can provide feature importance, the interpretability of the model may not be so intuitive as its complexity [Khandani, A. Kim, and Lo 2010].

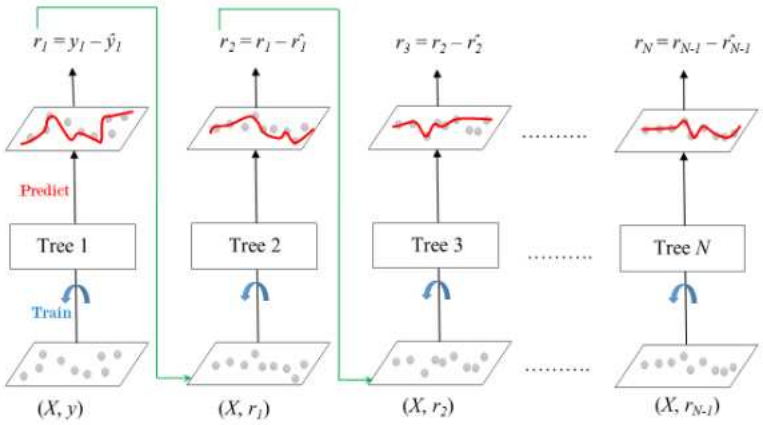


**Figure 3.3:** Gradient boosting implementation

- **Support Vector Machines**

  SVMs are a type of supervised learning model that separates data using the concept of finding an optimal hyperplane that has separated the various classes such as reliable/unreliable borrowers.

  Credit risk assessment can use SVMs in the classification of borrowers as per the probability of default by using both linear and nonlinear

kernels to adapt to complex datasets [Bazarbash 2019]. They are powerful in handling high-dimensional data and hence effective in cases where classes are not linearly separable.

Their ability to find a clear margin of separation makes them robust in several types of classification tasks. However, SVMs can be hard to interpret since the decision boundary may not give insight into the decision-making process. Apart from that, by nature, they do not natively provide any probability estimates on classifications, which may be critical in risk assessment [Frolov and Lavrentyeva 2019].

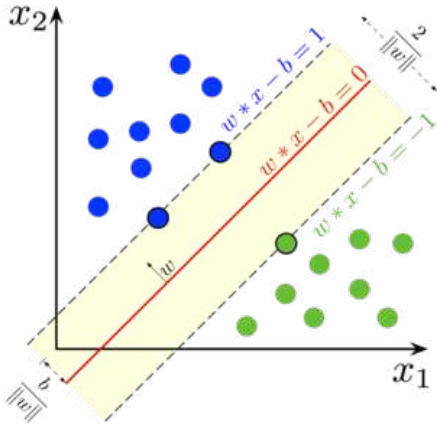


**Figure 3.4:** Support Vector Machine example

- **Neural Networks**

  Neural networks are collections of interconnected nodes or neurons, mimicking the human brain. They are designed to learn non-linear, complex relationships among input features.

  Through neural networks, large volumes of data in credit risk assessment will be elaborately analyzed for complex patterns across borrower behavior, payment histories, and economic variables to provide considerable insight into one's creditworthiness. By way of the many advantages coming their way, neural networks have been particularly cited for modeling complex relationships and interactions in data, applicable

to fraud detection and risk modeling [Bazarbash 2019]. They can also be further refined as more data becomes available.

The disadvantages of neural networks include, but are not limited to, the fact that the outcome of a neural network is largely considered "black box," wherein interpretation of internal workings is an issue. Lack of transparency can thus be an irritant to lenders who have to provide specific reasons for their decisions regarding outputs from neural networks. Training the neural network demands high computational resources and a deliberate, quite-easy adjustment of architecture and hyperparameters.

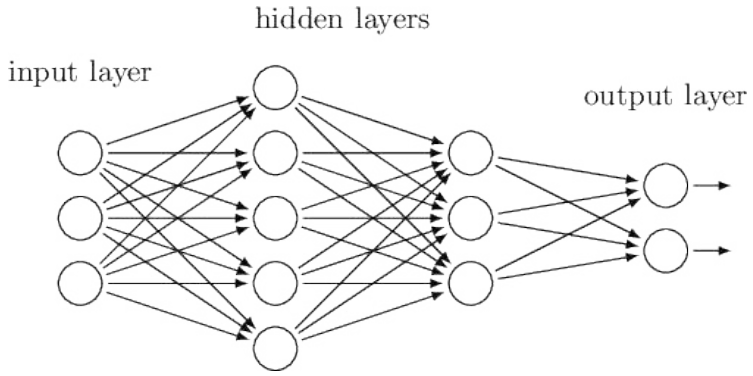


**Figure 3.5:** Neural Network layers

The inclusion in the credit risk monitoring process of such techniques as decision trees, random forests, gradient boosting, support vector machines, and neural networks thus offers certain considerable advantages: an improvement in predictive accuracy, in particular and a simple handling of complicated data [Thomas, Crook, and Edelman 2017]. Yet, all these models bring particular challenges, especially when there is a strong need for good interpretability along with the overfitting risks. A future balance between the advantages of machine learning and needs for transparency or ethical considerations should be the greatest concern to make sure that such advanced technologies are put to appropriate use.

### 3.1.2 Advantages and Disadvantages of Machine Learning Models

There are many benefits associated with machine learning models. These benefits make them quite appealing for many applications, more so in fields related to credit risk assessment. As in [Bazarbash 2019] , in fact, "the advantages associated with these mechanisms are trailed by challenges and limitations".

One of the strong suits of ML models is their capability to process large quantities of data, especially complex and unstructured data. Whereas in classical techniques, often structured data is available on pre-defined variables, an ML algorithm can easily assimilate data from transaction records, payment history, and even social media activities. The ability to leverage these kinds of diverse data points lets the models uncover numerous patterns and trends that might just pass through traditional analytical means. Therefore, it gives the lenders a much better insight into the behavioral pattern and financial status of the borrowers to make prudent decisions. Also, the ML algorithms tend to do remarkably well while unmasking hidden correlations and advanced patterns in data. These hidden correlations can improve the accuracy of credit risk predictions manifold. For example, a machine learning model may uncover a relationship between some aspect of an applicant's online behavior and their propensity to default, thereby allowing a lender to better assess their risk than they would have by evaluating traditional metrics in isolation. This can be quite important for improving the predictive power of models applied to forecast key metrics such as Loss Given Default or Probability of Default.

Another big plus in the use of machine learning is the fact that it relies on data from alternative sources. While processing information not from traditional sources but, for example, from social media activity or transaction patterns, the ML model generates a broader view about one's repayment

capability [Bazarbash 2019]. The integration of alternative data will be very useful especially for people with thin credit files, where the lender can make more equitable decisions with more indicators.

Of course, machine learning models come with their own particular set of challenges and potential pitfalls. At the top of these concerns is the possibility of biased decision-making. If biased against a certain direction, then the training data will keep the ML models free from any biases they will learn and amplify to build discriminatory practices in lending. This may extend to the very rejection or granting of unfavorable terms to access credit by certain demographic groups, which raises very serious questions about the morality of such decisions using automated lending systems.

Some of those algorithms, especially neural networks, can also be rather opaque and very complex. Because of this lack of explainability, it becomes hard for lenders to understand certain decisions taken by some models [Bazarbash 2019]. Such opacity has the effect of reduced trust in the system since lenders cannot explain or justify certain credit decisions either to the borrowers or to regulators and stakeholders.

Last but not least, some machine learning models appear to have specific problems catering to the current market conditions, which happen to change very fast. Quite often, the financial world experiences sudden turns in either economic ups and downs, changes in the regulatory environment, or completely unexpected events. If the ML model is not designed for such changes, then it is likely to fail and result in faulty risk assessments, which can be harmful to both lenders and borrowers.

While machine learning models unlock obvious benefits for such tools in terms of data capability, predictive accuracy, and alternative data, great risks and challenges come hand in hand. If these models are to be deployed responsibly and safely in the lending process, there is an overwhelming need to address issues related to bias, transparency, and adaptability. The full

capacity will thus be realized from a balance between exploiting the strengths of machine learning and mitigating its risks in credit risk assessment.

### 3.1.3 Comparison of Machine Learning and Traditional Econometrics

Machine learning models over traditional econometric methods do present some intimidating benefits when it comes to credit analysis [Bazarbash 2019], a field where such models have actually revolutionized the way lenders make decisions on the creditworthiness of a borrower. Among the most striking strengths of machine learning is the capability to process nontreeTraditional and nonlinear information. Whereas in traditional econometric models, assumptions of linearity often feature, together with a predefined set of variables, machine learning algorithms could handle complex data structures considerably better [Thomas, Crook, and Edelman 2017]. They are able to identify intricate correlations and interactions within the data that exist and thus enable understanding of borrower behavior in terms of detailed risk factors. Especially relevant for today's data-rich environment, all the sources of data (from transactional records to social media activity) can get amalgamated into insights regarding a borrower's creditworthiness.

It also considerably reduces information asymmetry in lending decisions. Because such models actually integrate a number of data sources, they finally create a superior profile of a borrower's financial health, capturing factors that may have been missed by traditional models [Van Gestel 2009]. This will also enable lenders to make more informed decisions and, hence, arrive at more realistic risk assessments and suitable lending strategies. Machine learning models show an improvement in performance regarding predictions for events such as LGD and other relevant metrics in credit risk assessment. They very often outperform traditional econometric models due to their ability to learn from large datasets and adopt patterns arising in the behavior

of borrowers.

The application of machine learning in credit analysis does not come without many challenges, ethical dilemmas, and perils. Most prominently, an issue that has been strongly highlighted can be stated as financial exclusion [Bazarbash 2019]. If machine learning models are either designed without due care or applied without due care, there may be a real risk that they will perpetuate existing biases in the data and, hence, make lending decisions that are discriminatory. Such biased algorithmic decisions might deny certain demographic groups much-needed access to credit on unfair grounds, which raises very serious ethical concerns about fairness and equity in lending.

Further, reliance on alternative data exposes the privacy and security of users. In other words, sensitive information on social media or personal spending habits could raise ethical issues regarding user privacy and data protection. Lenders have to navigate this ethics dilemma in data and balance respecting borrowers' privacy with useful intelligence from data [Thomas, Crook, and Edelman 2017].

Finally, structural changes to the market or economy can be challenging for machine learning models to manage. These models are great at recognizing patterns through experience, however when that experience suddenly shifts (for example, during times of economic downturn or a changed regulatory environment) these models can become somewhat ineffective [Bazarbash 2019]. This inability to quickly adapt to the new realities creates a significant risk given that in turbulent economic conditions, agility and responsiveness are often crucial.

In sum, while machine learning models offer powerful tools for credit analysis, improving predictive accuracy and reducing information asymmetry, they also carry inherent risks related to bias, privacy, and adaptability [Van Gestel 2009]. What is thus required is a judiciously balanced approach that tempers the strengths of machine learning with an intense awareness

of the challenges posed by the technology in the effective and responsible application in the financial sector.

While the application of machine learning to credit risk assessment is a great opportunity to increase access to credit and financial inclusion, the tool should be responsibly used and with awareness. In any case, with regard to the maximization of benefits coming from ML, it is necessary to adopt balanced approaches that provide guarantees regarding the level of transparency, continuous monitoring, and adjustment of models also in the light of external changes.

# Chapter 4

# Early Warning System

In this thesis we want to apply some machine learning techniques to implement a monitoring system that is capable of intercepting the signal of possible deterioration of some credit exposures of a bank. These systems are called "Early Warning System" and their purpose is to understand if a certain counterpart will have credit issues up to a certain time horizon [Kuritzkes and Schuermann 2006].

We want to test the performance of our algorithms on the target variable defined as the combination of:

- number of overdue days: the target is fixed as 31 days past due within 3 months;
- rating evaluation: the value of the rating associated with the position is increasing to "high" within 3 months.

The target variable can indeed be 1 if one of the conditions holds, otherwise it will be 0. We choose this combination as a dependent variable as it can give us a certain margin before the position will be considered as Past Due (that coincides with 90 days of overdues) and so the results can be used by the monitoring structure to adopt different types of strategies to stem the risk associated to the counterpart. The efficacy of the strategies is

higher if the offices have enough time to work on the potential problematics and to analyze the global counterpart risks, that can be hidden in different sources such as repeated overdues, an increase in the probability of default and issues with the balance sheet.

Watching the classical algorithms used to intercept such issues there are several problems that emerge [Bazarbash 2019]:

- The classical algorithms based on the presence or not of some triggers are not efficient: the main problematic regard the presence of a high percentage of false positives and this is related to the noise and the oscillation that the variables used to intercept the issues intrinsically have. In fact these triggers have only the state 1 when a certain level is pass, or 0 when the value is below that level, and this lead to over-estimating some signals or to not considering combinations of signals that can give us more information [Kuritzkes and Schuermann 2006].

- The model implemented has some optimization limit: we can consider some combination of events but the models based on trigger events are not capable of dealing with possible non-linearity between the variables.

- Due to these two problems they also affect the efficiency of the monitoring structure, since it has more and more counterparts to monitor (due to a low precision) and they are not capable of capturing early signals of credit issues (due to a low efficiency in intercepting dangerous counterparts).

The purpose of the thesis is to implement and train a random forest that is capable of discriminating between "bad" and "good" clients, using a set of information that is properly of different information areas of the counterpart and also some macroeconomic variables present in a certain period.

## 4.1 Methodology

Given the basic definitions, the required theoretical background, and the relevant studies in the field under consideration analysis, the following stage concerns the process by which the objectives of this research thesis will be achieved. Therefore, in this chapter the attention will be given to the framework of the applied methodology of the project, to completely understand the proposed methodology, and also to separate each step of the process carried out, such as selecting and analyzing and pre-processing data, testing different machine learning models and selecting different evaluation metrics [Castro 2019].

The methodology used follows an incremental approach, defining different steps by which we want to analyze in a more sophisticated way the data. The first step is to analyze and process all the information present in the dataset, giving a detailed description of all features contained, by underlying the importance of the target variable (in our case called **bad**).

The focus on the variables is put on by using different techniques, starting from detecting statistical behavior, moving on to obtaining some information with graphical techniques and finally finding hidden characteristics and possible correlations and detecting and managing the possible outliers or anomalies present in the dataset.

The step of pre-processing the dataset is important to get the most out of any model applied. When the data are ready we test different machine learning models to understand and see which of them give the best performance, based on a set of different measures [Castro 2019].

Those classification models are evaluated against different evaluation criteria, which may or may not be useful, depending on the needs of the individuals or institutions: depending on the goal that the financial institution wants to achieve, we can push these models to maximize some metrics with respect to others in order to satisfy specific needs.

As programming language we choose in a first moment **SAS** and **SQL** (used on **SAStudio**) to download the dataset and to perform some basic cleaning of the dataset, and then we continue with **Python** programming language to run the experiment in the Jupyter notebook environment, with the implementation of the following libraries, which are widely used in machine learning:

- pandas, numpy and sklearn libraries for data analysis and processing techniques;
- matplotlib and seaborn libraries for data visualization and graphical representation of the libraries;
- sklearn for importing evaluation metrics which are used to evaluate the algorithm's performance.

# Chapter 5

# Dataset description and Preprocessing

The dataset chosen is an aggregation of 6 months of observation, that starts from January 2024 to June 2024. The database contains a set of different information on the financial and credit situation of the client, which tries to give a complete description of all the informative areas of the client [Castro 2019]. We choose to implement the study considering only private clients such as physical persons, self-employees or co-headers, leaving out societies and companies that have different structures and different key indicators to prevent the default. In particular we have the following features:

- **Demographic information**, which gives some characteristics of the private clients such as profession, salary and properties;
- **Behavioural information** that are some variables describing the type of exposures that the counterpart has with respect to the financial institution;
- **Bank account area** that gives information about overdues, magnitude of granted, overdues or utilized credit or mean of movements on the bank account;

- **External information** that are scores associated with the counterpart given by CRIF (Risk central).
- **Macroeconomic information** that are variables such as inflation, interest rates, GDP and other factors.

The dataset is composed of numerical variables (34) and a categorical variable.

## 5.1 Feature analysis

These parts, as specified above, havw an important impact on the final outcome and evaluation of the machine learning models, in fact here the purpose is to describe the feature and try to understand which can be more useful for the analysis.

The dataset represents a crucial component of the analysis infrastructure, offering a snapshot of the real-world phenomena seeking to investigate. Furthermore, a crucial part of this section is the discussion related to the data pre-processing steps undertaken to refine and optimize the dataset, ensuring its optimal usability for the machine learning part.

The features that composed our dataset are listed below (labels are written in Italian):

- **Demographic module**
  - **Protesti**: this represents the presence or absence of protest associated with the position;
  - **Professione**: is the indication of the sector and the employment of the counterpart, aggregated into different buckets;
  - **Pregiudizievoli**: represent the presence or absenceof prejudicial associated with the position;

- **descr_flag_prop_imm_res**: description of the type of real estate property, 1 if owner, 0 if mortgage or rent and -1 indicates that the client is a cointestation, so the label is not applicable;
- **descr_flag_accr_stip**: description of the methodology of addebit on the bank account: 1 if direct, 0 if not and -1 if there's a co-header.

- **Behavioural module**

  - **ra_max_n_rate_scad_sm**: is the number of overdue installments in the last 6 months;
  - **ra_dur_residua_mean**: is the mean of the remaining duration period of the installments;
  - **flag_mr**: represent the presence of virtual marginal reports, 1 if present or 0 otherwise;
  - **IMV_PORTAFOGL_M**: is the value of the total portfolio at the end of the month;
  - **AL_nm3_sc**: is the number of overdue months in the previous 3 months.

- **Bank account module**

  - **cc_max_nd_nda_am**: is the total number of debit movements on the bank account over the sum of credit and debit movements in the last year;
  - **cc_mean_n_mov_a_am**: is the mean of the credit movements in the last year;
  - **cc_mean_n_mov_d_rat_tm**: is the mean of the installment debit movement in the last quarter;

- **cc_num_gg_sconf_max_am**: is the mean of the number of overdue days in the last year; this feature is, for the purpose of the research, truncated at 30 days of overdue;
- **cc_s_a_tm**: is the relation between the overdue and the granted amount given by the bank. The greater the number, the greater is the risk associated with the position sice it has more debt with the financial institution;
- **cc_u_a_m**: is the relation between the utilized credit and the granted amount given by the bank. The greater the number, the greater is the risk associated with the position since the counterpart is using a large amount of the granted credit by the financial institution.

- **External module**

  - **SCOREADV**: is a synthetic score that gives an overview of the financial and credit stability of the position. The source is the Risk Central, which has access to the overall financial situation of a counterpart, also with respect to other credit or financial institution. This organization is part of the Italian Bank.
  - **cc_imp_part_sosp**: it's a flag that is active whenever a position has some pending lot with a positive amount;
  - **d_autoliq**: it's a flag that is valued 1 if the counterpart has the presence of self-liquidating dealing;
  - **IXUTRE_CV_TM**: indicates the utilized over granted with respect to other financial institution during the previous trimester;
  - **cr_nasoff12**: it's a flag that is active whenever a position is reported as "suffering" by the Risk Central;

- **IRI_TM**: is the number of information requests from other financial institution regarding the counterpart. We consider the quarter fraction of the annual number;
- **cr_ixscto12**: it's a flag that is active whenever a position has an overdue reported by the Risk Central;
- **IXSCTOBIS_AM**: indicates the annualized overdue over granted with respect to other financial institution, considering only signature credits;
- **IXUTML_CV_AM**: indicates the utilized over granted with respect to other financial institution only for medium to long-term loans or credits;
- **naff_AM**: is the number of entities that are reported a certain position during the year;

- **Macroeconomic module**

  - **House_price**: is the variation of the house prices on a yearly basis computed every trimester;
  - **interest_1y**: is the interest rate yield curve in 1 year;
  - **interest_5y**: is the interest rate yield curve in 5 years;
  - **interest_10y**: is the interest rate yield curve in 10 years;
  - **interest_20y**: is the interest rate yield curve in 20 years;
  - **inflation_monthly_basis**: is the inflation rate computed monthly on the previous month;
  - **inflation_year_basis**: is the inflation rate computed monthly on the previous year;
  - **GDP**: is the GDP of Italy computed every trimester;

Now that all the features are defined the next step is to prepare the data for applying the machine learning algorithms. In the next section we will delve into the statistical and structural characteristics of the variables and we will produce some analysis and graphical representation of the features to describe their importance.

Moreover, the data processing chapter is essential to create a framework that describes in a detailed way all the steps to clarify the methodology to guarantee the reliability of similar analysis.

## 5.2 Data Preprocessing

In this section we want to present the variables and their distribution and, moreover, to compute some metrics and evaluations of the feature with respect to the target variable and within each other. We start by producing a general description of the dataset, with classical values such as the maximum, minimum, mean and standard deviation [Castro 2019].

These initial data give us some ideas on the distribution of the variables and if there are outliers or missing values.

| | cc_imp_part_sosp | descr_flag_prop_imm_res | protesti | descr_flag_accr_stip | cc_mean_n_mov_a_am |
|---|---|---|---|---|---|
| **count** | 1.838894e+06 | 1.838894e+06 | 1.838894e+06 | 1.838894e+06 | 1.838894e+06 |
| **mean** | 1.680358e-04 | -8.921558e-02 | 7.417502e-04 | -6.105028e-02 | 3.085123e+00 |
| **std** | 1.296178e-02 | 8.619038e-01 | 2.722500e-02 | 8.804975e-01 | 5.238536e+00 |
| **min** | 0.000000e+00 | -1.000000e+00 | 0.000000e+00 | -1.000000e+00 | 0.000000e+00 |
| **25%** | 0.000000e+00 | -1.000000e+00 | 0.000000e+00 | -1.000000e+00 | 1.333333e+00 |
| **50%** | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 2.333333e+00 |
| **75%** | 0.000000e+00 | 1.000000e+00 | 0.000000e+00 | 1.000000e+00 | 3.666667e+00 |
| **max** | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.000000e+00 | 1.491000e+03 |

**Figure 5.1:** General feature statistics

In the following part we treat the missing values of our dataset. This step has fundamental importance for the data integrity of our dataset: in fact missing values can lead to a bad fitting of the machine learning model

or to a bad evaluation on the validation/test set.

Moreover, credit monitoring models, both traditional models and machine learning algorithms, need complete and consistent data for training and validation. Addressing missing values is essential for accurately segmenting borrowers into risk categories. Incorrect categorization due to missing data can lead, as we said before, to inappropriate allocation of resources, since the structure that has to monitor the positions could be burdened by many practices and moreover this can lead to bad management and supervision [Castro 2019].

Innovative EWS should take into account the missing issues and manage this problem in order to mitigate model risk and to give a better and clearer evaluation of the position.

In our dataset we can clearly see that there are some features (as in Figure 5.2 in the fourth line) that contain missing values, here indicated with the numbers $-99999$ or $-99992$, which represent the "missing values" and the "value with 0 at the denominator".

| IMV_PORTAFOGL_M |
|---|
| 1.838894e+06 |
| -6.060697e+04 |
| 1.441105e+05 |
| -9.999900e+04 |
| -9.999900e+04 |
| -9.999900e+04 |
| 1.016750e+01 |
| 3.920871e+07 |

**Figure 5.2:** Missing values

Dealing with missing values requires introducing data imputation techniques.

## 5.3 Data Imputation with Missing Values

As discussed in the paper [Kwak and J. H. Kim 2017], there are various approaches for handling missing values in numeric data, depending on the nature and cause of the missingness. Missing data mechanisms are categorized into MCAR (Missing Completely at Random), MAR (Missing at Random), and MNAR (Not Missing at Random).

MCAR assumes that missingness is entirely random, meaning the likelihood of missing data is unrelated to both observed and unobserved variables. In contrast, MAR assumes that missingness depends on observed information, while MNAR implies that missing data depend on unobserved information, leading to potential biases in analysis if not handled appropriately.

Techniques for addressing missing values range from simpler methods, like listwise deletion (which excludes records with missing values but can significantly reduce dataset size and potentially introduce bias) to more complex imputation methods that approximate missing values to maintain dataset completeness, enabling deeper analysis [Castro 2019].

To handle missing values, several imputation techniques are employed, including single imputation methods (e.g., mean, median, and regression imputation) and k-Nearest Neighbors (kNN) imputation, as well as multiple imputation methods like Predictive Mean Matching and Bayesian Linear Regression [Kwak and J. H. Kim 2017].

In our dataset, features affected by missing values include **IMV_PORTAFOGL_M**, **cc_s_a_tm**, **cc_u_a_m**, **ra_max_n_rate_scad_sm**, and **ra_dur_residua_mean**.

For the specific case where values are represented as $-99992$ (notably in columns **cc_s_a_tm** and **cc_u_a_m**), we chose to replace these values with the mean or median of the feature. This is based on our understanding that these cases reflect overdue or actively used positions without an assigned credit amount. Using 0 would be inaccurate, as it would mix accounts with

regular payments or no overdue with accounts that have irregularities, potentially skewing analysis [Castro 2019].

For other missing values, we applied additional imputation techniques, using the median value in some cases based on related features.

For example, for **ra_max_n_rate_scad_sm**, we examine **cc_s_a_tm** and **cc_num_gg_sconf_max_am** and if either variable is not 0, the missing value is set to the median of the feature. Similarly, for **ra_dur_residua_mean**, we reference **cc_mean_n_mov_d_rat_tm** to check if installment debit movements are present.

Finally, we cleaned the dataset by eliminating rows with missing values that could not be reasonably imputed.

### 5.3.1 Feature selection

As we can see from the synthesis, some features are not informative for our purpose. Since we have to delete certain records due to the high proportion of missing values, several features have lost their significance [Castro 2019].

In the plot below (Figure 5.3), we can observe that these variables either have identical values throughout or have the majority of their values clustered within a narrow range.

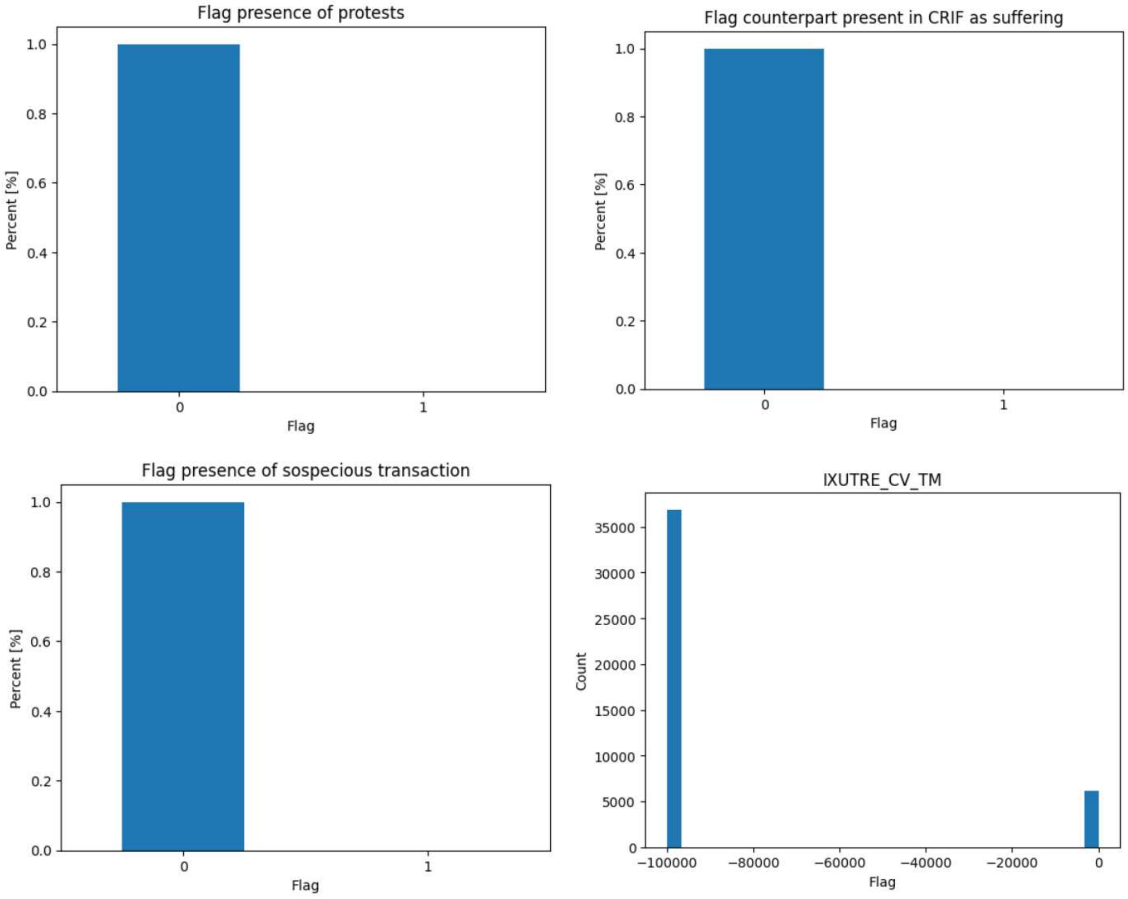


**Figure 5.3:** Almost constant variables

These features not only provide limited information on the "**bad**" probability but can also negatively impact the performance of machine learning algorithms. When splitting the dataset into training and test sets, these features could be inadvertently assigned more weight rather than being disregarded, which may bias the model.

So we exclude from the dataset the columns named **cr_nasoff12**, **IXUTRE_CV_TM**, **cc_imp_part_sosp** and **protesti** since the values, reported in Table 5.1, were:

| Values | cr_nasoff12 | cc_imp_part_sosp | protesti |
|---|---|---|---|
| 0 | 47959 | 47952 | 47938 |
| 1 | 6 | 13 | 27 |

**Table 5.1:** Unbalanced classes in the feature

Moreover we decided to exclude **IXUTRE_CV_TM** for the presence of a lot of missing values (36884, here indicated as $-99999$).

As we can clearly see, the majority of the values are concentrated in 0, and this doesn't give us useful information on the target variables. Moreover if we try to apply the split between test and training set, it can happen that all the records belong to the first or the second dataset and this can lead to a wrong training or to a poor generalization.

### 5.3.2 Outliers detection and treatment

Outlier detection and treatment are some of the major processes in data analysis because these extreme values disturb the statistical estimations, thereby producing misleading values [Kwak and J. H. Kim 2017]. Outliers are those elements of data that lie very much away from the usual range of distribution in a dataset. They may originate from a variety of sources, including human error in data entry, abnormal responses on the part of participants, or real but infrequent occurrences in the data. If left unchecked, outliers can bring in high bias, particularly in statistical measures such as the mean, standard deviation, or regression coefficients, which result in overestimation or underestimation of the values many times [Castro 2019].

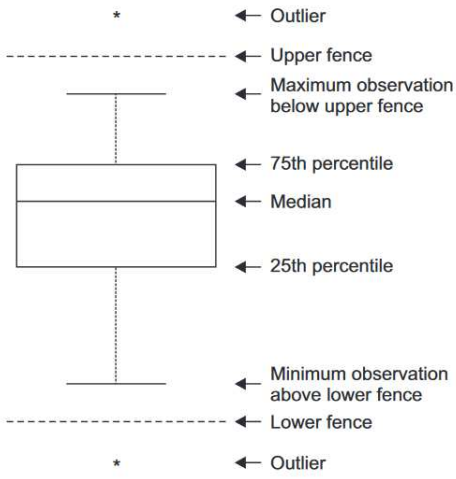


**Figure 5.4:** Outliers representation

Outlier detection is a sequential statistical process of methods for identifying these unusual values in a dataset. Among the common methods for outlier detection, one relies on the evaluation of the distance of the data point from the mean [Kwak and J. H. Kim 2017]. The data that falls outside a distance of three standard deviations from the mean are usually considered to be potential outliers. Because means and standard deviations are themselves sensitive to outliers, other methods such as the median and interquartile range are used in practice. Box plots, which graphically represent distributions of data, may also be used to identify outliers in terms of those points that fall outside the calculated "fences" of the IQR. Residuals and standardized residuals in regression analysis detect outliers based on their deviation from the predicted values, while more complex data call for support vector methods [Castro 2019].

After the detection of outliers, treatment involves methods of mitigating their effects without tampering with the integrity of the data. The following are the three main methods in handling outliers:

- Trimming: In this approach, the outliers are entirely removed from the data set. While trimming decreases the effect of extreme values

and dispersion, it also decreases sample size and, hence, might diminish the representativeness and power of any analysis. This process is easily carried out; however, it may not be proper because once outliers involve relevant information and not data entry errors, bias may appear.

- Winsorization: This method replaces the outlier values by pulling them towards the rest of the data range without actually removing the data. Winsorization can be performed either by replacing outliers with the highest or second-lowest value within the acceptable range of data, or by altering the weights of extreme values. This is useful in that, with this technique, not a single data point gets lost, yet again the influence of extreme values is restricted.

- Robust Estimation: This approach employs statistical models that are, by their very nature, less vulnerable to the effects generated by outliers. Robust estimation ensures a dependable result under conditions of outliers. For example, estimators that use the median and those techniques that are not dependent on extreme values render the results impervious to the impact of outliers. However, robust estimation procedures are usually computationally complex; hence, a higher degree of statistical knowledge is often needed.

By detecting and treating the data through such methods, the analysts are in a better position to control the possible biases caused by outliers. This is very crucial in trying to ensure that their statistical inferences are reliable and accurate, while avoiding all those errors in drawing conclusions from such data [Castro 2019]. A thoughtful treatment of outliers at the stage of research design guarantees that the results of a study are statistically valid to actually portray patterns in data, rather than artifacts brought about by unexamined extreme values.

In our dataset we detect some variables that have outliers, such that

**cc_mean_n_mov_a_am**, **cc_s_a_tm**, **cc_u_a_m** and **IMV_PORTAFOGL_M**.

As we can see there are three out of four features that had also problems with the missing values, and so we can firstly think to delete some of them, since they can lead to a wrong training.

Analyzing more in depth the variables we can see their distributions.

- **cc_mean_n_mov_a_am**: this feature represents the mean of the credit movement in one year and it's distributed as follows:

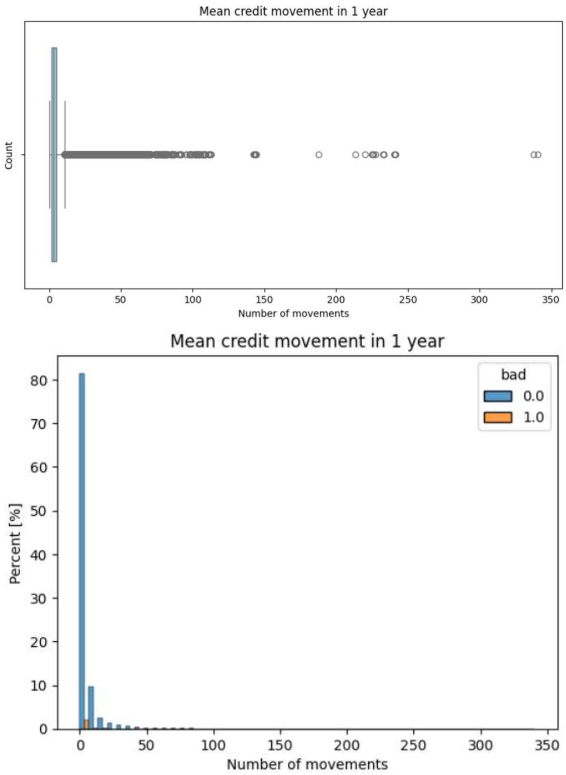


**Figure 5.5:** Distribution of the mean credit movement

We try to cut the tail of the distribution since it is quite uncommon to have more than 1 or 2 credit movements in one month (that are 12 - 24 in one year) and so we decided to eliminate the values of the variables that were above the 95% percentile. In Figure 5.6 we can see the final distribution:

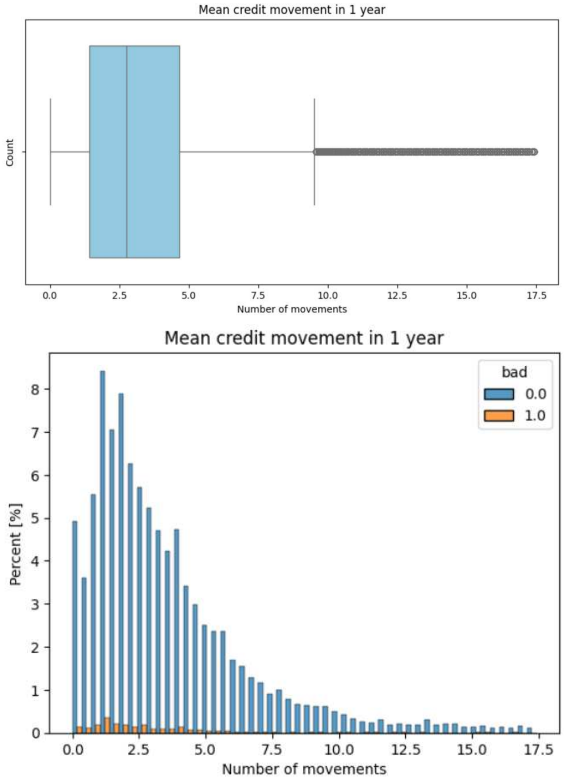


**Figure 5.6:** Distribution of the mean credit movement after processing

- **cc_s_a_tm**: this feature represents the mean of the overdue over the granted amount given by the financial institution (in %). The initial distribution is:

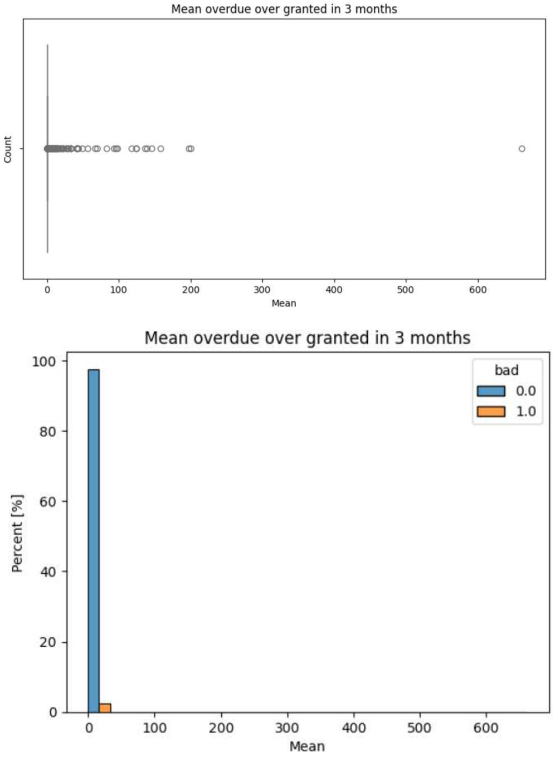


**Figure 5.7:** Distribution of the mean overdue over granted

Also in this case we try to cut the tail of the distribution since there are values that are extremely high (like the counterpart with 600 times the granted in overdues). In Figure 5.8 there is the final distribution:

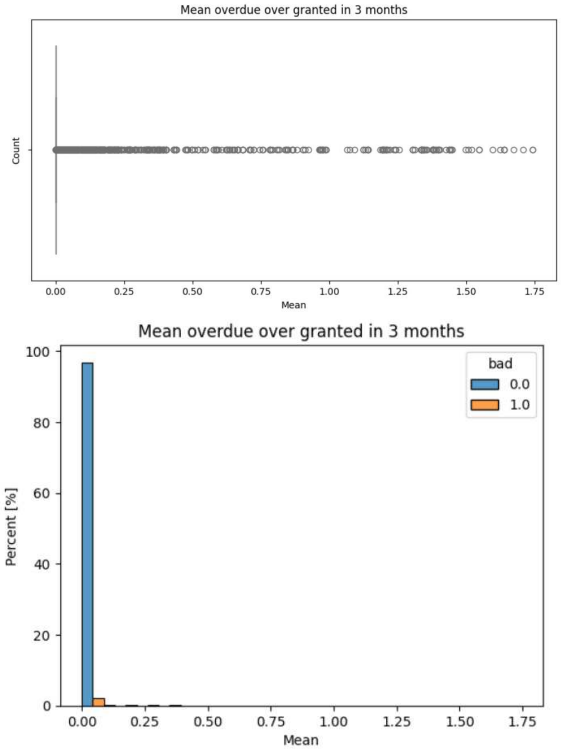


**Figure 5.8:** Distribution of the mean overdue over granted after processing

- **cc_u_a_m**: this feature represents the utilized credit over the granted amount given by the financial institution (in %). The initial distribution is:

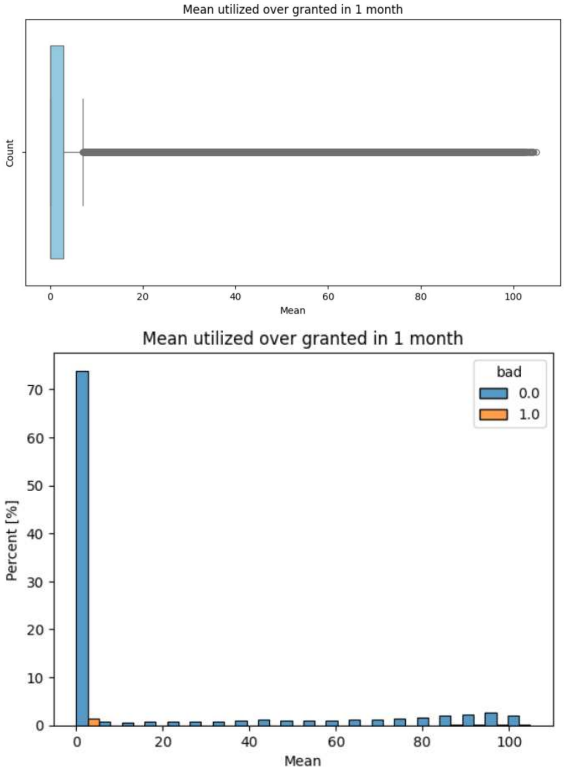


**Figure 5.9:** Distribution of the utilized over granted

As we can see in Figure 5.9 there are some outliers but, as the utilization of the credit granted indicates positions that can have possible future problems, we decided to let this variable be as it was in the dataset. In fact the bigger the ratio the bigger is the possible credit risk associated to the counterpart.

- **IMV_PORTAFOGL_M**: this feature represents the value of the total portfolio at the end of the month. Since the variance is really high we cannot cut too much the outliers, but we try to reduce the presence of extreme values. In fact we expect that a large amount of the population (since we consider private clients) has less than 50000 euros. The initial distribution is:

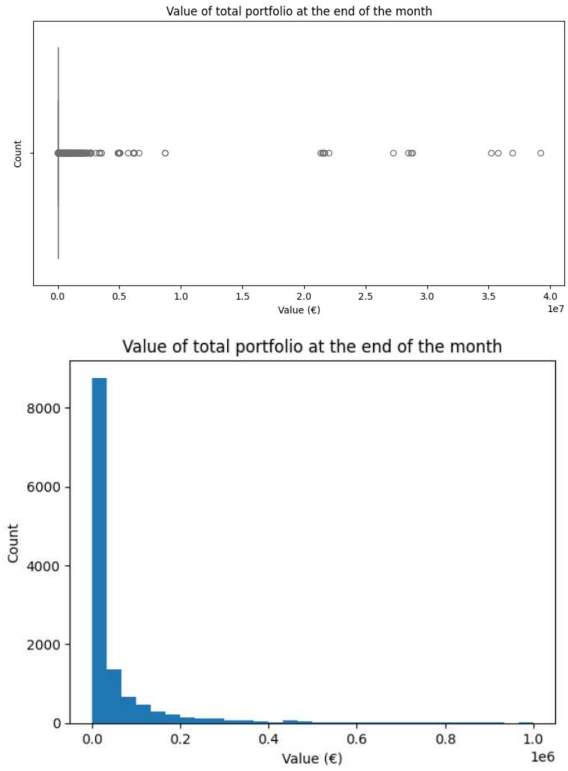


**Figure 5.10:** Distribution of the portfolio value

We decide to cut the extreme values of the tails, probably those that represent private clients with a certain heritage that aren't common and that can lead to wrong generalization and training in the models. In Figure 5.11 we plot the final distribution:

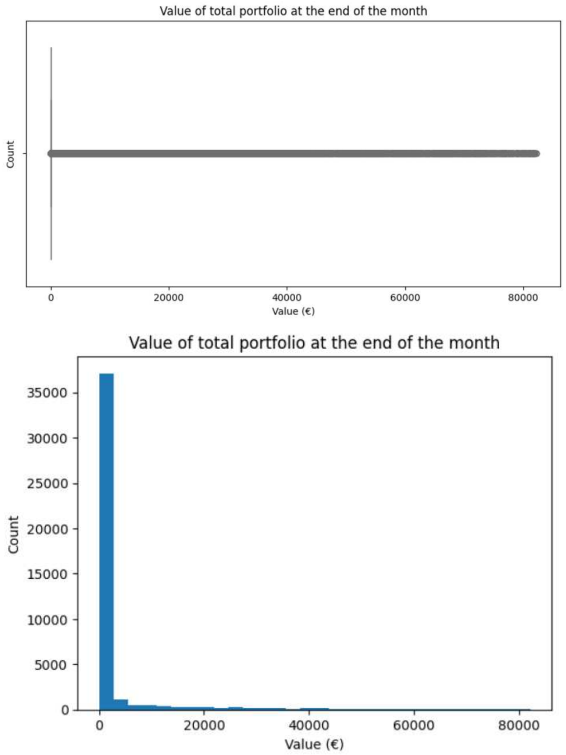


**Figure 5.11:** Distribution of the portfolio value after processing

### 5.3.3 Categorical feature analysis

The only categorical variable in our dataset is the profession, here called **Professione**. This variable is the result of the aggregation of more granular employment that we had aggregated in order to obtain fewer classes that were representative and correlated within the group.

For instance, students are more similar to housewives because they didn't have an income or a fixed salary every month, or moreover employees are very different from artisans or self-employed people but they have a similarity with police workers or firefighters.

As we can see from the figure, we found 7 different classes:

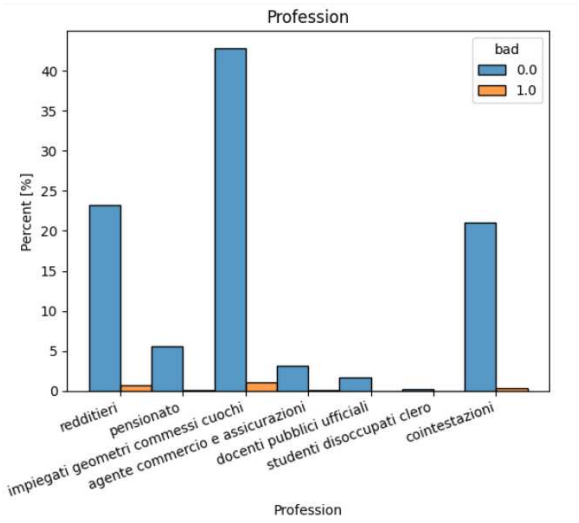


**Figure 5.12:** Feature "Profession"

The classes are not balanced since the majority is represented by self-employees and manufacturing, then we can find the people who have a fixed income (such as employees) and then we notice the presence of co-headers, who are basically bank accounts or loans shared by two people that are married or live in the same house.

From the plot we can see that there's not a clear division within the same class: in fact the portion of the bad counterparts and the good ones is the

same within the class.

### 5.3.4 Encoding and bucket creation

Encoding categorical variables is a vital step in preparing data for machine learning tasks. When dealing with categorical variables that are characterized by non-numeric values such as text or categories, it becomes necessary to transform them into a numerical format for compatibility with machine learning algorithms [Castro 2019]. Various categorical encoding techniques are available, each presenting its unique set of advantages and drawbacks. Some of the most used categorical encoding methods are:

- One-hot encoding: it's one of the most important encoding tecniques, that transforms a categorical variable into an n-dimensional array with 1s and 0s;
- Ordinal encoding: transform the categorical variable into ordered integer, giving a sort of "order importance" to the variables;
- Count encoding: count the presence of the variable in the dataset and replace the label with this number;
- Target encoding: replace the categorical variable with the mean of the target variable for each category.

In our case we choose to perform the one-hot encoding technique since it was more suitable for nominal categorical variables, where the categories have no inherent order or relationship. The idea behind one-hot encoding is to represent each category as a binary vector.

For each category in the **Professione** categorical column, a new binary column is created. The binary column will have a value of 1 if the class is present, else it will be zero. So we start from here:

| Feature name | Category |
|---|---|
| professione | agente commercio e assicurazioni |
| | cointestazioni |
| | docenti pubblici ufficiali |
| | impiegati geometri commessi cuochi |
| | pensionato |
| | redditieri |
| | studenti disoccupati clero |

**Table 5.2:** Profession category before encoding

and then we end up with 7 columns encoded with the label **professione2_"Category"**.

| Column name | Array value |
|---|---|
| professione2_agente commercio e assicurazioni | [1,0,0,0,0,0,0] |
| professione2_cointestazioni | [0,1,0,0,0,0,0] |
| professione2_docenti pubblici ufficiali | [0,0,1,0,0,0,0] |
| professione2_impiegati geometri commessi cuochi | [0,0,0,1,0,0,0] |
| professione2_pensionato | [0,0,0,0,1,0,0] |
| professione2_redditieri | [0,0,0,0,0,1,0] |
| professione2_studenti disoccupati clero | [0,0,0,0,0,0,1] |

**Table 5.3:** Profession category after encoding

The same treatment was done for the variables **descr_flag_accr_stip** and **descr_flag_prop_imm_res** which has values of 1 or 0 for the presence or absence of the indicator and there is another value (-1) that represents when the value cannot be computed (since, for example, if we have a co-header we can't say if they have both a salary or if they have both a property).

Here in Figure 5.13 are the plots also for these two features, which were converted into a 3-dimensional array (Tables 5.4).

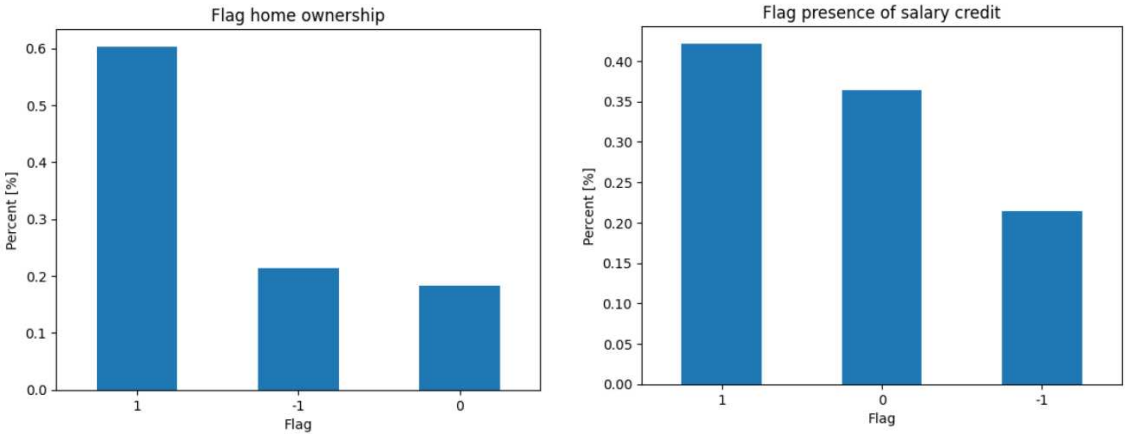


**Figure 5.13:** Home flag and Salary flag before encoding

| Column name | Array value |
|---|---|
| descr_flag_accr_stip_1 | [1,0,0] |
| descr_flag_accr_stip_0 | [0,1,0] |
| descr_flag_accr_stip_-1 | [0,0,1] |
| descr_flag_prop_imm_res_1 | [1,0,0] |
| descr_flag_prop_imm_res_0 | [0,1,0] |
| descr_flag_prop_imm_res_-1 | [0,0,1] |

**Table 5.4:** Home flag and Salary flag after encoding

One of the other techniques we adopt in our analysis is the creation of a bucket in order to discretize variables and simplify the learning phase for our algorithms.

The variables we decided to process are **IXSCTOBIS_AM**, **IXUTML_CV_AM**, **naff_AM**, **cc_max_nd_nda_am_2**, **cc_mean_n_mov_d_rat_tm_2**, **cc_u_a_m_2**.

Data binning is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values that fall into a given small interval, a bin, are replaced by a value representative of that interval, often a central value (mean or median).

Statistical data binning is a way to group numbers of more-or-less continuous values into a smaller number of "bins". For example in our case, we want to group the different counterpart exposures into representative bins (for example, grouping every 20.000 euro). It can also be used in multivariate statistics, binning in several dimensions at once.

## 5.4 Correlation analysis

This forms the backbone of any default prediction analysis, hence shedding light on the variables that are most strongly associated with the event of default. These can then be used by any financial institution as predictor variables in credit risk models, hence providing more realistic predictions of default.

Most of the regulatory requirements in finance force the institutions to assess and manage credit risk in an efficient manner. In this regard, the application of correlation analysis in practices related to risk management will contribute to the meeting of such regulatory requirements [Castro 2019].

Correlation analysis finds widespread application in stress testing. A stress test is normally carried out by a financial institution to identify how its portfolios would perform under adverse economic conditions. Correlation

analysis allows modeling of scenarios where several variables change together. This offers important insights into the potential loss when economies go into recessions. These correlations can also help in a much more effective use of capital: a lender can distribute its capital in relation to different loan segments' risk, based on how those segments correlate with one another.

With this, institutions can give focused strategies in the areas of risk management peculiar to each segment their loan portfolio falls into [Castro 2019]. They can have different terms for borrowers falling in segments with lower or higher risk profiles, interest rates, or mitigation measures.

Generally, a correlation analysis between variables will result in improvement in the quality of risk assessment, diversification of loan portfolios, enhancement of models for predicting defaults, and assurance of conformance to regulatory requirements.

In Figure 5.14 we can see the correlation of the feature with each other and with respect to the target variable, which is called **bad**.

On the diagonal of this matrix, we can see the correlation of each feature to itself, always equal to 1. The matrix, for every cell, stores the correlation between the variable represented on the vertical axis and that represented on the horizontal axis. Colors represent the direction and strength of the correlation: positive correlations go from yellow to red, while negative correlations go from yellow to blue.

This matrix is useful because it gives an overview of the linear relationship between features. When the correlations are very high or very low, this will help to detect variables able to provide redundant information. Such variables can be removed from the dataset in order to avoid possible problems during training by machine learning algorithms.

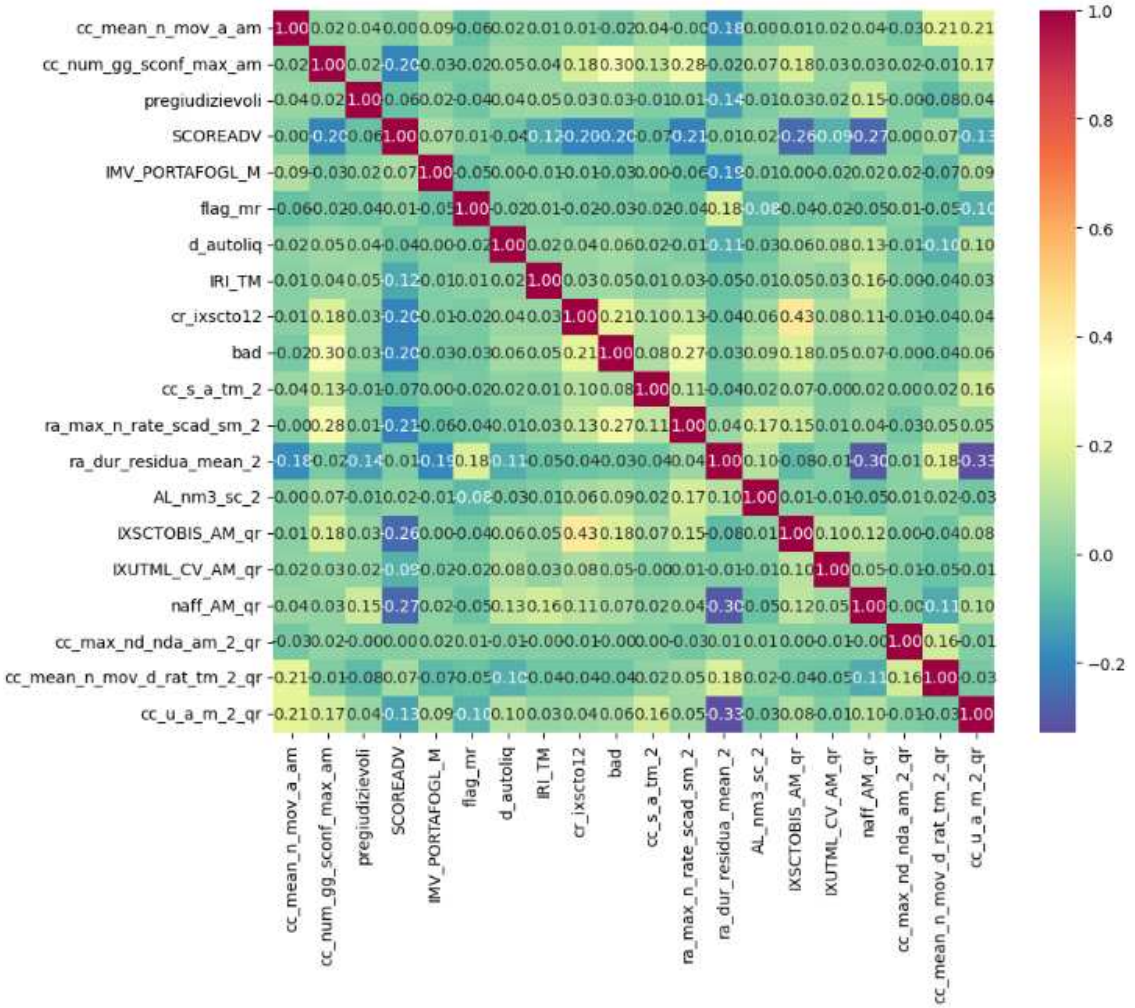


**Figure 5.14:** Correlation analysis

On the other hand, the matrix enables finding patterns across different features, as shown in Figure 5.14. For example, it is observed that the number of days past due is highly related to the number of overdue installments. This makes much sense because if a borrower has a lot of days past due, chances are he/she missed one or more installments from previous months. Some features are also highly correlated with the target variable. As an instance, if the borrower already has the problem of payment, then the probability of having a rating increased risk or days overdue over 30 is

higher.

The matrix also indicates that several of the "problematic variables" in describing the credit risk, at least one such as the number of days overdue, the existence of overdue installments, and being placed in a central risk register, are inversely related to the external score mentioned here as SCOREADV. That is to say, the "good" customer will generally lead to a higher external score, while the appearance of credit issues or past delinquencies lowers this score.

## 5.5 Target variable

The last but more important analysis is the analysis of the target variable, that is the true outcome that the machine learning algorithms should predict.

The distribution of the target **bad** is shown below:

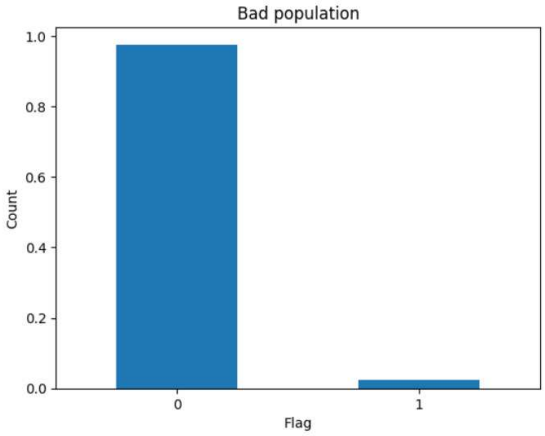


**Figure 5.15:** Dependent variable

The data were highly imbalanced, something expected in any financial institution. Only a small fraction of their customers ever have credit problems. Such imbalance can make the problem challenging for algorithms of credit risk because it may imply difficulty in correctly training the model and assessing its performance. Besides, conventional metrics of performance such as accuracy are meaningless in this context since they do not reveal the class imbalance problem.

For example, assume a model classifies all clients in our dataset as "good"; the metric of accuracy would still show about $\sim 98\%$, since defaults are less than 3% of the total observations. This high value from accuracy may look impressive, but it is misleading because the model failed in identifying the minority class, which is crucial in credit risk evaluation.

To handle this problem, techniques must be employed that balance the dataset in such a way that more "bad" observations are created for use in training by the algorithm. In this thesis, we have chosen to employ the

SMOTE method due to its performance as an efficient balancing technique.

SMOTE stands for Synthetic Minority Oversampling Technique and it is a technique developed to deal with the problem of imbalanced classification when the instances of the minority class are so few that a model cannot learn from them [Nitesh 2002]. Traditional methods of oversampling, such as simply duplicating the existing examples from the minority class, do balance the dataset but don't convey any new information. SMOTE does this more effectively by generating synthetic examples for the minority class in order to improve the generalization capability of a model [Castro 2019].

The process of SMOTE works like this: the algorithm picks up an example from the minority class randomly. It determines a given number, usually five, of nearest neighbors from this selected example, still within the same class. From these neighbors, it picks out one randomly from which to develop a synthetic example by creating a point along the line segment connecting the two examples [Castro 2019]. The new point is obtained by interpolation between the original example and its chosen neighbor using a random factor to place the synthetic point anywhere along that line segment in space. It ensures that the data points so generated will be plausible, for they will resemble the existing examples in feature space.

This process can be repeated an arbitrary number of times in order to generate sufficient synthetic examples that balance the dataset. Often, in practice, SMOTE accompanies random undersampling of the majority class [Nitesh 2002]. This helps avoid the potential problem of over-diversification of the sample space by focusing the model on the most important decision boundaries. Performance from this strategy is often better than from using undersampling alone.

Here we reported a simple example in order to visualize the concept: in Figure 5.16 we see the original distribution of the dataset for the target variable with respect to two features.

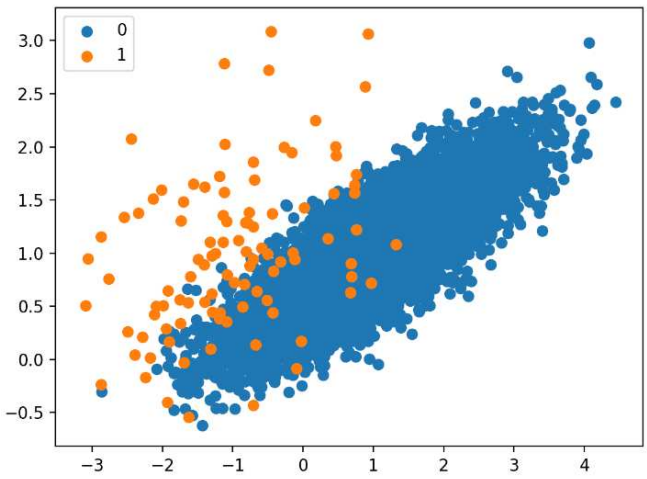


**Figure 5.16:** Initial dataset with minority class

Applying the over-sampling technique of the minority class with SMOTE we end up with a dataset that is distributed as follows:

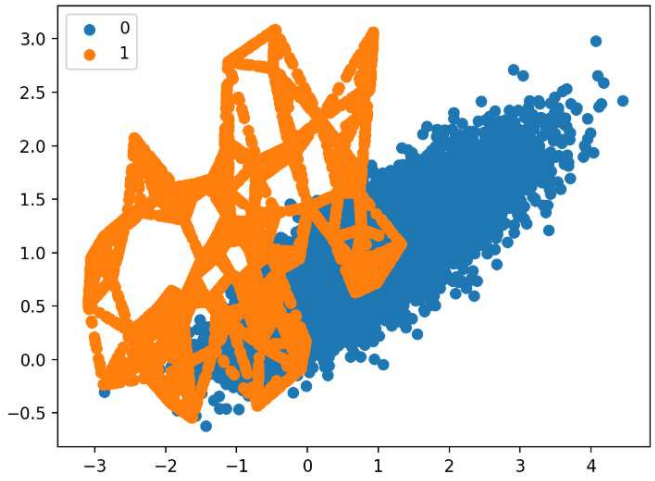


**Figure 5.17:** Final dataset after the application of SMOTE

This works well because the approach increases the decision boundaries around the minority class, thereby allowing the classifier to learn better patterns and structure from the minority class. One potential disadvantage of using SMOTE is that it generates synthetic examples without considering the majority class and if the two classes happen to overlap in feature space, then

this may create ambiguous examples that will puzzle the classifier [Nitesh 2002].

Another analysis we want to present is the feature importance, which is computed as the correlation between the features and the target variable to understand the capacity of some of them to discriminate between good and bad custumers. Here in Figure 5.18 are the results:

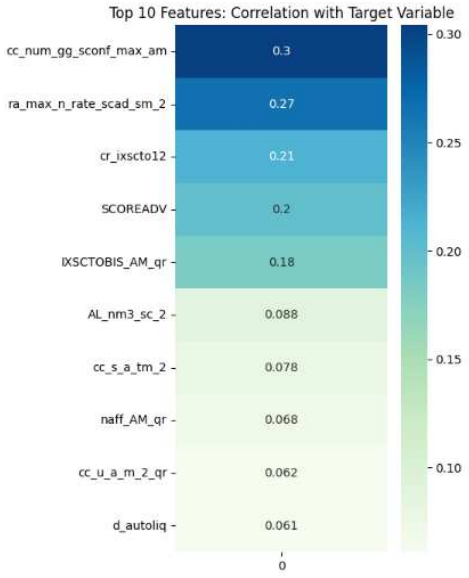


**Figure 5.18:** Correlation with the target variable

Here we can highlight the fact that there are only a few features that present a strong correlation with the target variable, and this gives us the idea to investigate in a deeper way the possible combinations and non-linearity between the variables.

# Chapter 6

# Algorithms implementation

The problem of understanding the behavior of the combination of different signals and how they can affect the capacity of a client to pay the credit given by the bank led us to investigate the thematics using machine learning techniques.

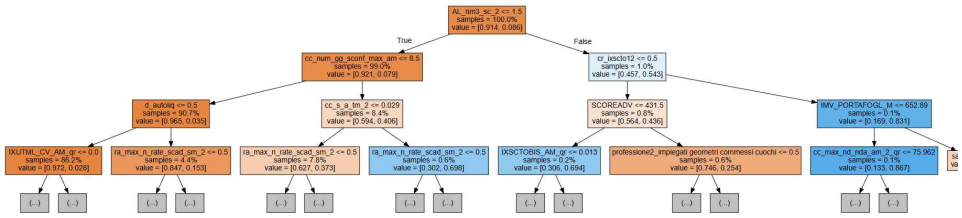


**Figure 6.1:** Tree example in the Trained Random Forest

In Figure 6.1 we can see the representation of one of the decision trees of the trained Random Forest. In this case we see that some of the features are taken into account in the tree and the initial sample is split into two subgroups if the condition in the box is true or not for the observation. The process is repeated until we reach the leaves of the tree in which every observation is given a label, corresponding to the class 1 or 0.

We decided to stop the tree representation at the 3rd node for a better visualization but the tree is set to have a deeper construction, fixed by the parameter **max_depth**.

We choose to implement a Random Forest using firstly the dataset as it was after the preprocessing and then we implement the SMOTE techniques to see if there is an improvement in the performance. We moreover decide to implement also a Gradient boosting algorithm to see if there is an improvement in the predictions.

We decided to proceed with a Random Forest due to its ability to enhance predictive accuracy and mitigate overfitting of decision tree models. In fact one of the most important difficulties of decision trees is the poor generalization due to the adaptability of the algorithm to the training set.

The Random Forest model is robust and reliable for credit risk default prediction, given that its ensemble learning approach aggregates a large number of decision trees' predictions [Biau and Scornet 2016]. This aggregation results in less overfitting, hence generalizing better. By exploiting complex and nonlinear relationships among various credit risk factors, Random Forest strengthens predictive accuracy, with feature importance scores enabling financial institutions to identify critical risk factors and make better-informed decisions when lending [Partnoy 2017].

Despite the above benefits, there are still disadvantages of the Random Forest. In fact as more trees are added to a forest, the interpretability of the model decreases, making it harder to provide explanations for single predictions. Feature importance scores are intuitive to understand, but understanding the reasoning for a given output can be complicated [Castro 2019]. Another issue is that Random Forests tend to be computationally much heavier than single decision trees, and this may be a problem if operations of credit risk assessment are large. In most cases, this trade-off between complexity and predictive power will make the Random Forest a very useful resource when interpretability is not very critical, and accuracy is more important.

The model performs very well for credit risk default prediction, and for

several reasons. First, in ensemble learning, each tree is developed on different random subsets of data and features, hence the overfitting issue can be minimized while capturing a wide array of risk patterns and interactions. Second, it can handle high-dimensional and complex data with efficiency while supporting both categorical and continuous features [Castro 2019]. Its adaptability extends further to non-linear relationships of features with the likelihood of default, hence boosting accuracy and precision. Also, the effect of averaging predictions over trees has a smoothing effect on the impact of noisy or outlier features in any one tree.

Gradient Boosting is a concept of supervised machine learning, part of the ensemble learning family that can be applied to classification or regression problems. Random Forest and Gradient Boosting algorithms are ensemble methods because these algorithms create a final model by combining the outputs of many individual models.

Gradient Boosting trains a sequence of models by giving higher weights to those examples whose prediction error is high so that the next model in the sequence can be trained. It iteratively trains a new model with the goal of minimizing a loss function similarly to how ANNs learn the optimal weights.

In the Gradient Boosting algorithm, weak learners are fitted one after another, for which their predictions get contrasted with the reality. The difference between prediction and reality gives the error rate of the model. Now, this error rate would be used to calculate the gradient, which is the partial derivative of the loss function. The gradient suggests how the parameters in the model need to be modified toward reducing the error in subsequent iterations during training.

Unlike neural networks, which only train one model on a loss function, Gradient Boosting combines the predictions of several models to achieve better overall performance [Castro 2019]. Some of the hyperparameters are

shared with algorithms such as Random Forest: for example, the number of trees and the maximum depth. However, some of the unique hyperparameters in Gradient Boosting come from neural network-like hyperparameters such as learning rate and loss function.

## 6.1 Tuning hyperparameter

First of all, we implement the Random Forest and the Gradient Boosting algorithms with some default parameters in order to see how the performance was when choosing standard values.

Then first of all we try to set the weights to give to the output of the model: in this way we can decide to give a precise weight to the outcome 1 ("bad" client) and another to the opposite class. This is useful because we train the model to give more importance to the class we want to predict better.

Using Python, we call the GridSearchCV method that, given an initial grid of points (in this case the weights to pass to our model), applies the Crosss Validation technique to understand which is the best set of parameters that suites our model better. Here is the code in Python:

As we can see we set some default parameters and then we search in the grid of points that are the best for our case. We found that the best parameters were:

| **Parameters** | **Value** |
|---|---|
| $w_0$ | 0.19 |
| $w_1$ | 0.81 |

**Table 6.1:** Weight tuned

where $w_0$ is the weight associated to the class 0 and $w_1$ with class 1.

For the rest of the study we keep this parameters fixed. The next step is

```
weights_1 = np.linspace(0.1,0.3,20)

param_grid_1 = {'class_weight': [{0:x, 1:1.0-x} for x in weights_1]}

# Create a random forest classifier
rf =  RandomForestClassifier(n_estimators=300, max_depth=40,
                             min_samples_leaf = 2, random_state=60)

# Use random search to find the best hyperparameters
grid_search_1 = GridSearchCV(rf,
                           param_grid = param_grid_1,
                           cv=5)

# Fit the random search object to the data
grid_search_1.fit(X_train, y_train)

best_rf_1 = grid_search_1.best_estimator_
```

**Figure 6.2:** Weight hyperparameter tuning

to set the remaining parameters of the Random Forest which are precisely:

- **max_depth**: The max_depth parameter decides how tall the trees in the forest are allowed to get. This is a very important parameter regarding model performance, since it controls overfitting. While higher tree depth can sometimes lead to higher accuracy up to a point, beyond that a model typically has increased chances of overfitting and learning noise rather than patterns in data. This setting should therefore be set so that overfitting is avoided. Default is None, which means that the tree will keep unfolding its nodes until all leaves are pure or contain fewer samples than specified by the min_samples_split parameter.

- **min_samples_split**: The parameter of min_samples_split, by default, is the minimum number of observations in a node for splitting to be allowed. By default, it is set to 2, meaning that a node can be split if it contains more than two samples and if the node is not already pure. However, that default is going to frequently cause trees to be overgrown, splitting down into nodes until they are only purely

homogeneous. Such a tree will be oversized and tends to overfit.

- **min_samples_leaf**: The min_samples_leaf defines the minimum number of samples that can exist in a leaf node following a split. This is done to prevent the tree from creating very small, fragmented leaf nodes.

- **n_estimators**: The n_estimators controls the number of decision trees that constitute the Random Forest. Indeed, increasing the number of trees lets the model improve in performance by reducing the variance, but this will not result in overfitting. On the other hand, using too many trees increases computational time and complexity dramatically. In scikit-learn, by default, this value is set to 100, and it generally provides a good trade-off between performance and efficiency.

We set, as in Figure 6.2, a grid of hyperparameters where the algorithms have to search the best combination.

```
n_estimators = range(100,400,50)
max_depth = range(10,100,10)
min_samples_split = range(2,6)

param_grid = {'n_estimators': n_estimators,
              'max_depth': max_depth,
              'min_samples_split' : min_samples_split}

# Create a random forest classifier
rf = RandomForestClassifier(class_weight=class_weight,
                            random_state=60, min_samples_leaf = 2)

# Use random search to find the best hyperparameters
grid2_search = GridSearchCV(rf,
                            param_grid = param_grid,
                            cv=5)
```

**Figure 6.3:** Random Forest hyperparameters tuning

And the final set of best parameters is reported in the Table 6.2:

| Parameters | Value |
|---|---|
| **max_depth** | 20 |
| **min_samples_split** | 2 |
| **n_estimators** | 300 |

**Table 6.2:** Random Forest hyperparameters tuned

The last step is the implementation of the Gradient Boosting. As we already trained a Random Forest we consider the parameter found before thanks to the grid search algorithms.

Firstly we test the algorithms to understand the default performance and then we tune the remaining parameters, which are in this case:

- **Learning Rate**: Another important hyperparameter is the learning rate, which determines the magnitude of the contribution to the final forecast by every single tree. Gradient Boosting starts with an initial estimate and then iteratively improves it with further outputs of other trees. These updates essentially get scaled by the learning rate. The preference is for lower learning rates, however it tends to result in a more robust model that prevents overfitting to the individual trees and increases generalization. However, lower learning rates also require more trees to be able to capture the relationships in data; that is, increasing computational cost.

- **Loss**: the loss function is the metric that the algorithm seeks to minimize at each step. Default loss functions work most of the time. It's better to use another option when there's a clear understanding of the effect it will have on the model.

So in order to also tune the Gradient Boosting hyperparameters we implement another grid search for the Learning Rate parameter. Here in the Figure 6.4 Python code:

```
learning_rate = np.linspace(0.1,0.3,20)

param_grid_gmb = {'learning_rate': learning_rate}

# Create a random forest classifier
gmb =  GradientBoostingClassifier(n_estimators=300, max_depth=20,
                            min_samples_leaf = 2, random_state=60)

# Use random search to find the best hyperparameters
grid_search_gmb = GridSearchCV(gmb,
                           param_grid = param_grid_gmb,
                           cv=5)
```

**Figure 6.4:** Gradient Boosting hyperparameter tuning

and the best parameter found is:

| Parameters | Value |
|---|---|
| **learning_rate** | 0.28 |

**Table 6.3:** Learning rate tuned

## 6.2 Introduction of Macroeconomic variables

In this section we want to introduce in our dataset also some macroeconomic variables.

Our purpose is to enlange the set of variables and to give more information on the general financial and credit situation of Italy and the Italian economy, testing if this information gives us more details and if the combination of a personal and general credit situation results in a more precise and accurate prediction in monitoring credit issues [Huangfu et al. 2024].

We start introducing the variable in the dataset and, as the features are on a monthly or trimestral basis, they will be the same for the counterparts present in the same determined period of time.

First of all we compute the correlation analysis and we see that these

economic informations are really correlated with each other, as reported in Figure 6.5.

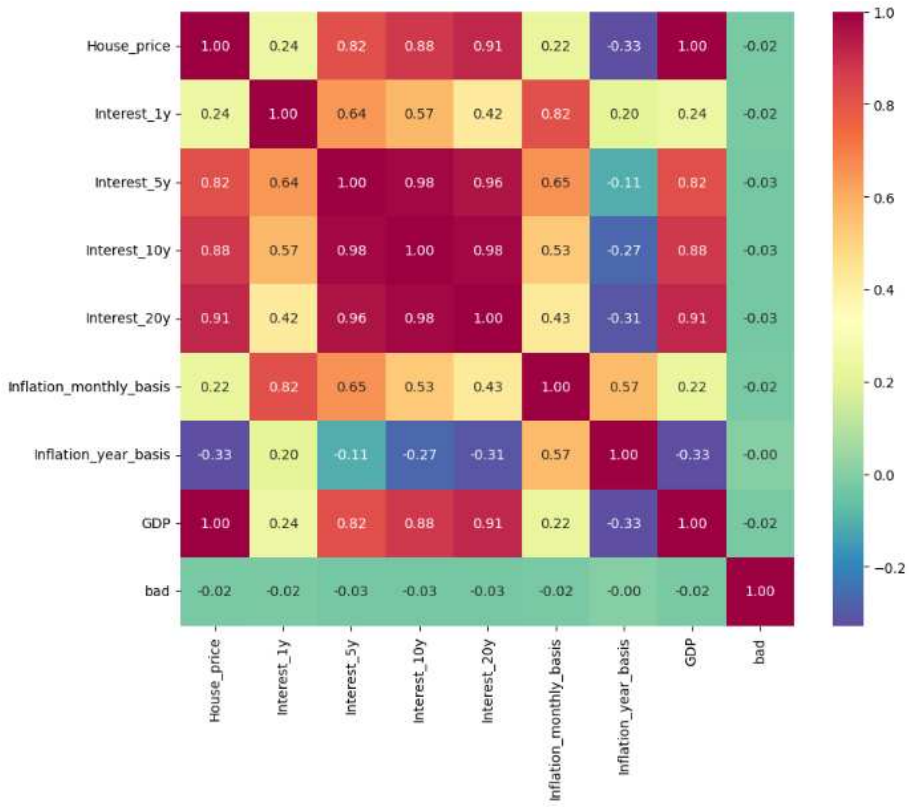


**Figure 6.5:** Correlation matrix for macroeconomic variables

Moreover, if we plot the first 10 most correlated variables with respect to the target variable, we can see that the list doesn't change, as shown in Figure 6.6 .

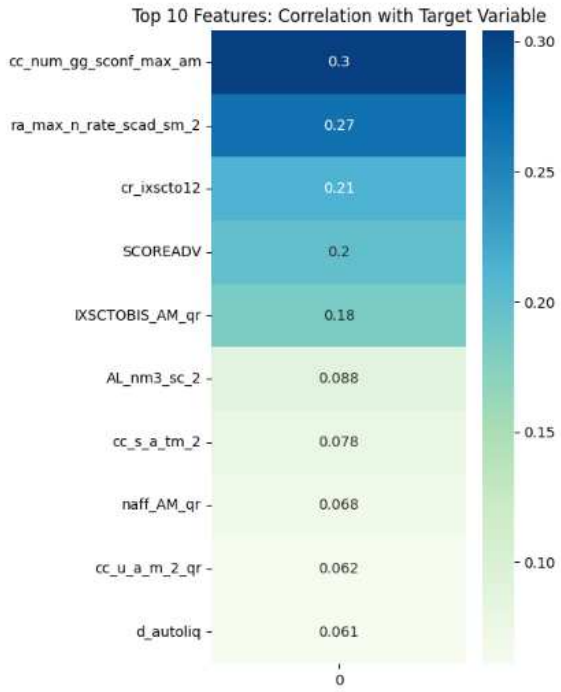


**Figure 6.6:** Correlation with target variable

As we have already tuned the algorithms and found the right parameters, we test the two ML algorithms on the enriched dataset to see if the results change [Huangfu et al. 2024]. In the next section we report all the results and the metrics we choose to implement for the evaluation of the performance.

# Chapter 7

# Results

The last step of the study is to evaluate the performance of the algorithms and see if they can be useful for future application in the credit monitoring process.

Firstly, we have to choose which of the various metrics we want to compute and which are more useful for our purpose. As we said before during the study, metrics like accuracy can lead to a wrong evaluation of the final performance of the algorithms since the dataset (and also in a real-life situation) is totally unbalanced due to the presence of a lot of "good" counterparts and few "bad" ones, which however are the most interesting for the monitoring structure of a bank.

The best way to visualize the majority of them is through the confusion matrix, which is a double-entry matrix in which we see the possible combination of true label and predicted label:

Moreover in the next steps we will call the 4 combinations as TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative).

Here we present a set of metrics used to evaluate the performance and the formula used to compute the score.

- **Accuracy**: This is the metric that gives an overall feeling of correct-

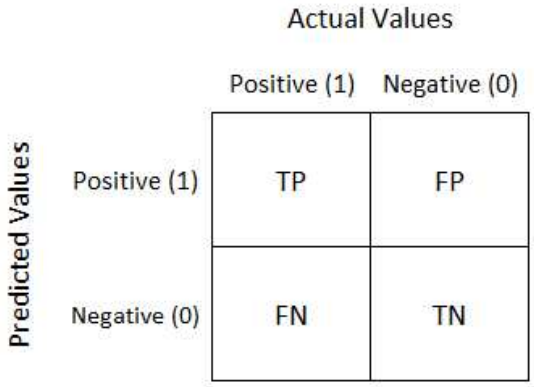


**Figure 7.1:** Confusion matrix example

ness on the forecast of a model. Accuracy can be found using the ratio between correct predictions and the total number of predictions given by the model.

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

- **Precision** : The precision or exactitude of the positives, say, tells what percent of the positive predictions out of all the positive predictions were correct. It is within the context of binary classification when one wants to avoid false alarms or positive misclassifications.

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

- **Recall**: This is the number of all actual positives that the model correctly predicts. It is high if the positive cases that were recalled are large.

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

- **F1 Score**: The F1 score is a measure of a model's performance based on both precision and recall into a single metric using their harmonic

mean. It provides a balanced evaluation of a model's performance, especially when there is an uneven class distribution.

$$Fscore = \frac{2 * Precision * Recall}{Precision + Recall}$$

- **ROC-AUC**: ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. To understand we can look at Figure 7.2

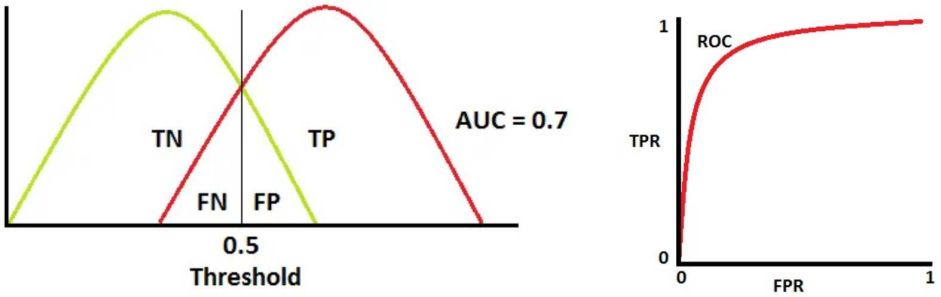


**Figure 7.2:** ROC-AUC curve

- **Matthews correlation coefficient**: MCC is a measure of the quality of binary classifications, which ranges from -1 to +1. A score of +1 represents a perfect prediction, 0 an average random prediction, and -1 an inverse prediction. The MCC is particularly useful when the classes are imbalanced, as it accounts for the imbalance in the calculation.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

First we report the results of the first Random Forest algorithm without using any under- or oversampling and without tuning the parameters. We want to maintain a good depth and number of trees (set to 40 and 300). We call this step **rf_no_tuned**:

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_no_tuned** | 0.98 | 0.78 | 0.12 | 0.20 | 0.56 | 0.30 |

**Table 7.1:** Results for standard RF

As discussed before, we notice that although the accuracy is really high the model has a good precision but a really low recall, which means it's not capable of intercepting bad clients from the population. The other metrics tell us the same behavior for this preliminarly model.

Moving forward we consider the same random forest but trained using the database in which we have applied over-sampling of the minority class. We expect that since the number of "bad" examples is more, the algorithm may improve his capability of prediction. This step is called **rf_SMOTE_no_tuned**:

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_SMOTE_no_tuned** | 0.98 | 0.97 | 0.79 | 0.87 | 0.89 | 0.87 |

**Table 7.2:** Results for RF using SMOTE technique

In the next step we perform the tuning of the weight class parameter and then, after verifying the performance, we perform a second tuning of the other Random Forest parameters. The results we obtain are reported in Table 7.3, with the names **rf_SMOTE_weight_tuned** and **rf_SMOTE_full_tuned**.

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_SMOTE_weight_tuned** | 0.98 | 0.94 | 0.84 | 0.89 | 0.92 | 0.88 |
| **rf_SMOTE_full_tuned** | 0.98 | 0.95 | 0.83 | 0.88 | 0.92 | 0.88 |

**Table 7.3:** Results for RF using SMOTE and tuned parameters

The last step before adding the macroeconomic features is to train the

Gradient Boosting classifier and tune the parameter for the learning rate. Here are the results for the standard XGBoost and the tuned one:

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **xgb_SMOTE_no_tuned** | 0.98 | 0.94 | 0.85 | 0.90 | 0.92 | 0.89 |
| **xgb_SMOTE_full_tuned** | 0.98 | 0.95 | 0.84 | 0.89 | 0.93 | 0.88 |

**Table 7.4:** Results for XGBoost with tuned parameters

Moving forward, we test the algorithms with all the tuned hyperparameters with the dataset containing the macroeconomic variables for the 6 months considered. Here are the results obtained with the Random Forest and the Gradient Boosting:

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_SMOTE_macro_tuned** | 0.98 | 0.94 | 0.82 | 0.88 | 0.91 | 0.87 |
| **xgb_SMOTE_macro_tuned** | 0.98 | 0.94 | 0.83 | 0.88 | 0.91 | 0.88 |

**Table 7.5:** Results for XGBoost and RF with the introduction of macroeconomic variables

So summarizing all the tables, we can see the behavior of the algorithms with all the improvements we decided to test in this thesis.

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_no_tuned** | 0.98 | 0.78 | 0.12 | 0.20 | 0.56 | 0.30 |
| **rf_SMOTE_no_tuned** | 0.98 | 0.97 | 0.79 | 0.87 | 0.89 | 0.87 |
| **rf_SMOTE_weight_tuned** | 0.98 | 0.94 | 0.84 | 0.89 | 0.92 | 0.88 |
| **rf_SMOTE_full_tuned** | 0.98 | 0.95 | 0.83 | 0.88 | 0.92 | 0.88 |
| **xgb_SMOTE_no_tuned** | 0.98 | 0.94 | 0.85 | 0.90 | 0.92 | 0.89 |
| **xgb_SMOTE_full_tuned** | 0.98 | 0.95 | 0.84 | 0.89 | 0.93 | 0.88 |
| **rf_SMOTE_macro_tuned** | 0.98 | 0.94 | 0.82 | 0.88 | 0.91 | 0.87 |
| **xgb_SMOTE_macro_tuned** | 0.98 | 0.94 | 0.83 | 0.88 | 0.91 | 0.88 |

**Table 7.6:** Results summary

As a last analysis, we want to implement a simpler model in order to make the output more intuitive and interpretable , thinking in a monitoring framework where the employee has to take decisions over the synthetic results given by the model. In this case having a smaller set of variables can lead to a simpler assessment and to efficient decisions.

In order to do this, we plot the most important features, selecting only the ones that are bigger than a certain level (here set at $2,5\%$) and the results are reported in Figure 7.3.

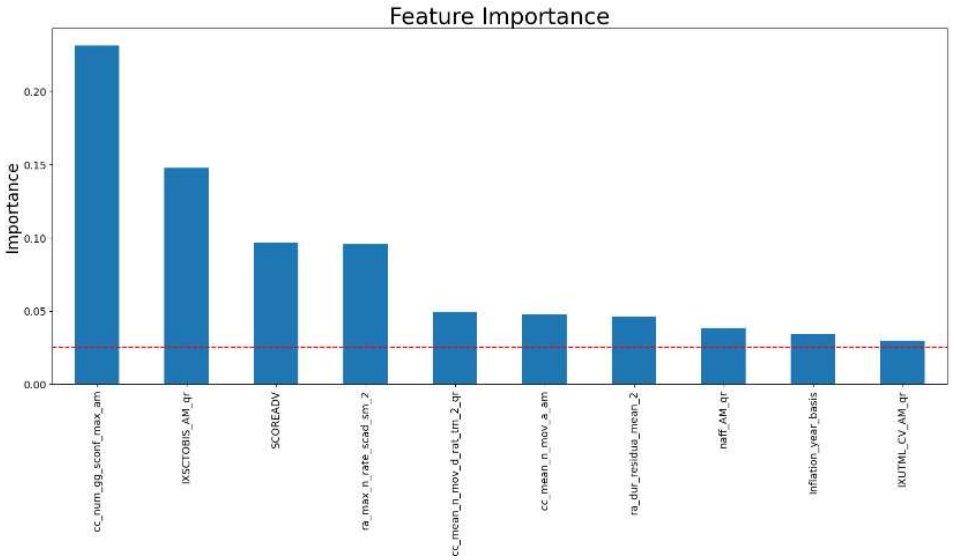


**Figure 7.3:** Feature importance

Using then the tuned random forest, we train the model and we test the results on the test set. The results are reported in 7.7.

| Model | Accuracy | Precision | Recall | F1 score | ROC-AUC | MCC |
|---|---|---|---|---|---|---|
| **rf_fs_tuned** | 0.97 | 0.87 | 0.81 | 0.84 | 0.90 | 0.82 |

**Table 7.7:** Results for feature selected Random Forest

As we can clearly see, the results are slightly worse compared to the previous models, but the performance is however high. In a financial framework, the decision between the models should be done considering not only the statistical and performance part but also more technical and logistic needs, such as the interpretability and the simplicity for the users.

We report in the next section all the comments regarding the previous results.

## 7.1 Conclusion

The purpose of this study was to investigate whether the use of Machine Learning (ML) and Artificial Intelligence (AI) could enhance the performance of the monitoring system of a financial institution, as well as to analyze the advantages and disadvantages of changing the assessment methodology.

The results, summarized in Table 7.6 and Table 7.7, indicate significant improvements. Specifically, the precision of the models is notably high, demonstrating that the algorithms effectively identify "bad" clients and additionally the recall values are medium-high, signifying that the Warning System successfully detects a substantial proportion of problematic counterparts before they default.

Overall, considering other performance metrics, it can be concluded that ML delivers excellent results, particularly in handling numerous variables and accounting for complex, non-linear relationships among features. However, a notable challenge arises after implementing such detection systems: designing efficient techniques for monitoring the structure. ML algorithms often lack interpretability, as they do not provide explicit reasons for labeling a client as "bad." Instead, labels are assigned based on complex relationships among features.

For example, as illustrated in Figure 6.1, tracing the decision-making path within a decision tree is complex. This challenge is compounded when using ensemble models like Random Forest, where decisions are aggregated from multiple trees, further complicating interpretability.

Lastly, when macroeconomic variables were added to the dataset, the overall performance remained similar to that of the original dataset. This may be attributed to:

- macroeconomic variables often have effect over the long period: here we

consider only 6 months and obviously we can't see significant variability in the interest rate or in house prices. Probably considering a loger period of time the results would be different;

- we consider only macroeconomic fctor at the time of observation: probably, for instance considering the interest rate, adding information of past values or values at the signature of the instalments would have more impact on the final outcome.

- we choose a limitated set of variables, which can be improved and tested with other indicators more related with credits.

For certain, we can conclude that with the increasing complexity of credit regulations, models such as Random Forest and Gradient Boosting offer significant improvements for the monitoring departments of financial institutions. These models are highly effective in early identification of potential problems, enabling institutions to take proactive measures to mitigate risks or reduce exposure when defaults occur.

Finally, the findings emphasize the importance of adaptability in model development, recognizing the dynamic nature of financial markets and the changing profiles of borrowers and clients. The identification of optimal features and the fine-tuning of model parameters are critical in achieving higher predictive accuracy and enhancing system performance.

This research provides valuable insights into the field of credit monitoring, supported by firsthand experience during my internship. It also underscores the ongoing necessity for continuous refinement and adaptation in the application of machine learning models to ensure their effectiveness in evolving contexts.

Looking ahead to the future of credit monitoring, it is clear that Early Warning Systems are a promising solution. However, their efficacy can be further improved through enhancements in interpretability and the quality

of input data. Collaboration among various departments within a financial institution is crucial to define a clear path and successfully implement these systems.

# Bibliography


Altman, Edward and Gabriele Sabato (2007). "Modelling credit risk for SMEs: Evidence from the US market". In: *Abacus* 43.3, pp. 332–357.

Barbagallo, Carmelo (2018). "Il sistema bancario italiano: situazione e prospettive". In: *Associazione per lo Sviluppo degli Studi di Banca e Borsa, Università Cattolica del Sacro Cuore di Milano, Bologna ( https://www. bancaditalia. it/pubblicazioni/interventi-vari/int-var-2018/Barbagallo-20180324. pdf, Downloaded: 15/03/2019).*

"Basel committee on banking supervision" (2011). In: *Principles for Sound Liquidity Risk Management and Supervision (September 2008).*

Bazarbash, Majid (2019). *Fintech in financial inclusion: machine learning applications in assessing credit risk.* International Monetary Fund.

Beerbaum, Dirk and Sammar Ahmad (2015). "Credit risk according to IFRS 9: Significant increase in credit risk and implications for financial institutions". In: *Sammar, Credit Risk According to IFRS* 9.

Bernanke, Ben (1993). "Credit in the Macroeconomy". In: *Quarterly Review-Federal Reserve Bank of New York* 18, pp. 50–50.

Biau, Gérard and Erwan Scornet (2016). "A random forest guided tour". In: *Test* 25, pp. 197–227.

Brown, Ken and Peter Moles (2014). "Credit risk management". In: *K. Brown & P. Moles, Credit Risk Management* 16.

Castro, Carlos (2019). "Moody's Analytics Advisors". PhD thesis. Nova School of Business and Economics.

Frolov, Daniil Petrovich and Anna Victorovna Lavrentyeva (2019). "Digital economics of foreign trade activities in action: institutional capacities and limitations of complex ecosystem regulation". In: *1st International Scientific Conference" Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth"(MTDE 2019)*. Atlantis Press, pp. 187–191.

Giordano, Luca and Antonio Lopes (2008). "Dimensione, localizzazione ed assetto giuridico nell 'analisi dell'efficienza del sistema bancario italiano". In: *Quaderno n 07/2008*.

"Global financial development report 2019/2020: Bank regulation and supervision a decade after the global financial crisis" (2019). In: *The World Bank*.

González-Aguado, Carlos and Javier Suarez (2015). "Interest rates and credit risk". In: *Journal of Money, Credit and Banking* 47.2-3, pp. 445–480.

Al-Gunaid, Mohammed, Irina Salygina, Maxim Shcherbakov, Vladislav Trubitsin, and Peter Groumpos (2021). "Forecasting potential yields under uncertainty using fuzzy cognitive maps". In: *Agriculture & Food Security* 10.1, p. 32.

Huangfu, Yubin, Haibo Yu, Zuoji Dong, and Yingman Wang (2024). "Research on the Risk Spillover among the Real Economy, Real Estate Market, and Financial System: Evidence from China". In: *Land* 13.6, p. 890.

Khandani, Amir, Adlar Kim, and Andrew Lo (2010). "Consumer credit-risk models via machine-learning algorithms". In: *Journal of Banking & Finance* 34.11, pp. 2767–2787.

Kuritzkes, Andrew and Til Schuermann (2006). "What we know, don't know and can't know about bank risks: A view from the trenches". In: *Wharton Financial Institutions Center Working Paper*.

Kwak, Sang Kyu and Jong Hae Kim (2017). "Statistical data preparation: management of missing values and outliers". In: *Korean journal of anesthesiology* 70.4, pp. 407–411.

Lamarre, Eric, Kate Smaje, and Rodney Zemmel (2023). *Rewired: the McKinsey Guide to Outcompeting in the Age of Digital and AI*. John Wiley & Sons.

Löppönen, Oula (2022). *Critical review on the European bank regulators' definition of default for forborne bank loans*.

Matz, Filip and Yuxiang Luo (2021). *Explaining Automated Decisions in Practice: Insights from the Swedish Credit Scoring Industry*.

Nikolopoulos, Konstantinos and Andreas Tsalas (2017). "Non-performing loans: A review of the literature and the international experience". In: *Non-performing loans and resolving private sector insolvency: Experiences from the EU periphery and the case of Greece*, pp. 47–68.

Nitesh, Chawla (2002). "SMOTE: synthetic minority over-sampling technique". In: *J Artif Intell Res* 16.1, p. 321.

Ongena, Steven and David Smith (2001). "The duration of bank relationships". In: *Journal of financial economics* 61.3, pp. 449–475.

Packer, Frank and Haibin Zhu (2012). "Loan loss provisioning practices of Asian banks". In: *Bank for International Settlements*.

Partnoy, Frank (2017). "What's (still) wrong with credit ratings". In: *Wash. L. Rev.* 92, p. 1407.

Perry, George, William Fellner, Robert Gordon, James Duesenberry, Robert Hall, Christopher Sims, William Nordhaus, Robin Marris, Thomas Juster, and John Shoven (1980). "Inflation in theory and practice". In: *Brookings Papers on Economic Activity* 1980.1, pp. 207–260.

Rokach, Lior and Oded Maimon (2005). "Decision trees". In: *Data mining and knowledge discovery handbook*, pp. 165–192.

Stulz, René (2001). "Does financial structure matter for economic growth? A corporate finance perspective". In: *Financial Structure and Economic Growth: A Cross-Country Comparison of Banks, Markets, and Development*, pp. 143–188.

Tajik, Mohammad, Saeideh Aliakbari, Thaana Ghalia, and Sepideh Kaffash (2015). "House prices and credit risk: Evidence from the United States". In: *Economic Modelling* 51, pp. 123–135.

Taylor, Alan (2015). "Credit, financial stability, and the macroeconomy". In: *Annu. Rev. Econ.* 7.1, pp. 309–339.

Thomas, Lyn, Jonathan Crook, and David Edelman (2017). *Credit scoring and its applications*. SIAM.

Van Gestel, Tim (2009). *Credit risk management, Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital*. Oxford University Press.

Vousinas, Georgios (2015). "Supervision of financial institutions: The transition from Basel I to Basel III. A critical appraisal of the newly established regulatory framework". In: *Journal of Financial Regulation and Compliance* 23.4, pp. 383–402.

# Ringraziamenti

Ringrazio il mio relatore Martino Grasselli per avermi seguito durante il percorso di studi universitario e la stesura finale della tesi di laurea, essendo sempre presente per chiarire i miei dubbi e rispondere alle mie richieste.

Ringrazio inoltre tutto l'ufficio di Performance Management Bonis, per avermi accolto durante lo stage e avermi dato la possibilità di approfondire le mie conoscenze nel settore dei crediti bancari. Ringrazio anche i miei responsabili Virginia e Stefano per avermi dato questa opportunità e Pietro per essere stato il mio "mentore" fin dai primi giorni.

Voglio inoltre ringraziare i miei genitori per avermi sostenuto durante il percorso e avermi dato i mezzi per affrontare il percorso universitario, mio fratello Mattia e i miei nonni e zii per essere stati sempre presenti e pronti a darmi un aiuto qualora mi servisse. Grazie per avermi dato tutta la serenità e la tranquillità che una famiglia può dare.

Un ringraziamento di cuore a tutti i miei compagni di palestra con cui mi sono allenato per anni (purtroppo un po' meno in questo periodo) e soprattutto al mio maestro Michele, capace di far emergere quella determinazione e quella sicurezza che mi contraddistingue ora e che mi ha aiutato fin da piccolo.

Infine voglio ringraziare tutti i miei amici, da quelli di più lunga data a quelli che ho avuto la fortuna di incontrare durante gli studi universitari. Grazie per aver reso più leggero questo periodo, per avermi aiutato nei momenti no e per aver condiviso i momenti più belli, che sia stato a distanza

con un messaggio o che sia stato di persona.

Un ringraziamento speciale va a Paola, che più di tutti mi è stata vicino. Senza di te sicuramente non sarebbe stata la stessa cosa questo periodo, grazie per farmi vedere il lato positivo delle cose ed esserci sempre nel momento del bisogno, grazie per correggermi quando serve e per aver fatto uscire il lato più nascosto di me. Sei quella persona che rende speciale ogni momento.