Ca' Foscari
University
of Venice

Master's Degree programme

in Finance

Final Thesis

Bankruptcy prediction via machine learning

Supervisor
Ch. Prof. Emanuele Aliverti

Graduand
Giacomo Beggio
Matriculation Number 869097

Academic Year
2021 / 2022

Abstract

Bankruptcy prediction is the problem of detecting financial distress in firms, which could potentially lead to their bankruptcy. A good prediction method can allow the company's stakeholders to take action in order to improve the business' financial health or limit their economic losses. Given its relevance, this problem has been analyzed since the 1930s, and a plethora of prediction models have been proposed, starting from univariate statistical models to more complex, multivariate approaches (like the famous Altman Z-Score). However, although these models performed well in the context that they were applied, their predictive power decreased dramatically when used in different scenarios, making them unreliable. Since the 1990s and with the beginning of the "Big Data Era", machine learning models have proved to be the superior choice for bankruptcy analysis, since they are more versatile and offer much better predictions. This study demonstrates the predictive power of machine learning models based on a dataset of more than 6000 Taiwanese firms and 95 financial ratios. Three models have been used: the Logistic Regression, The LASSO Regression and the Random Forest, which, after being tested and evaluated, proved their effectiveness in predicting bankruptcy.

Contents

Introduction

Bankruptcy prediction is the problem of predicting bankruptcy and various forms of financial distress. The event of bankruptcy has crucial importance for the firm's life: the firm ceases to exist, causes economic losses to its stakeholder and, depending on its economic relevance, can seriously damage the economy. In the past, there have been several economic downturns (caused by financial crises, pandemics and other major disruptive events) which caused the bankruptcy of many firms. Given the relevance of the topic, since the 1930s academics have tried to develop various types of prediction models which allow firms to recognize early warning signals so that they can fix them before the inevitable happens. From the first univariate models to more complex multivariate ones, there have been various approaches to solve this problem. In general, classic "compact" models like the Altman Z-Score (which paved the way in this field) proved to be rather simple ways to give accurate predictions in the context that they were developed. However, their simplicity is also their biggest weakness, as they are not reliable methods in contexts outside of those in which they were developed. Arguably their use of financial ratios was restricted to a very limited number which simply could not analyse the "whole picture". Since the 1990s, the advent of machine learning in various fields of study brought researchers to consider them for the problem of bankruptcy prediction, proving to be the superior choice nowadays.

This thesis supports the use of machine learning algorithms to predict bankruptcy through the analysis of a dataset of more than 6000 firms and 95 financial indicators. In the first chapter, the phenomenon of bankruptcy is analyzed from an economic point of view. This regards the functioning of the legal procedure, the advantages and disadvantages, the main indicators to assess the financial status of the firm and recent trends. In the second chapter, the importance of bankruptcy prediction models is highlighted. First, three classic models are described and critiqued due to their unreliability in modern times. Then, the motivations regarding the importance of machine learning algorithms are presented, followed by a general description of the analysis which will be conducted. In the third chapter, the whole statistical background is described, starting from the categorization of the problem and its relative features, and then going more into depth about the models fitted and the performance metrics used to evaluate them. The last chapter

consists of the analysis and the results that come with it. The section starts with a visualisation of data and information about the dataset (such as the type of variables included, their correlation and the distribution of the observations among the two classes). In the last part, data modelling is conducted with different techniques and models to account for the specific characteristics of the data. Last but not least, the superior model in terms of predictive accuracy is declared and conclusions are drawn.

1 The legal and financial framework of bankruptcy

1.1 Introduction

Bankruptcy is a legal proceeding initiated when a business is not able to repay debts or obligations. This process can begin on behalf of the debtor (that is the legal subject who will suffer from bankruptcy) or by one or more creditors, however, this is less common. This procedure essentially consists of the measurement and evaluation of the debtor's assets which will then be used to repay the liabilities. The filing is presented to the relevant bankruptcy court and is assisted by a specialized lawyer. The term 'bankruptcy' must not be confused with that of 'insolvency' because they are different (even if similar). First of all, bankruptcy is a legal process or court order, whereas insolvency is a state of financial distress happening when the value of total liabilities exceeds that of total assets (according to the International Revenue Service). Second, bankruptcy is a type of insolvency but it is not the only one. Third, bankruptcy is not the only 'solution' to a state of insolvency, in fact, there are some others such as the liquidation of assets or the debt restructuring of the firm. Additionally, depending on the country in question, bankruptcy may apply only to some legal subjects while insolvency is a more generalized term. For example, in the UK bankruptcy only applies to individuals and sole traders with unlimited liability, while insolvency applies to businesses as well.

1.2 Bankruptcy fraud

Bankruptcy is a legal procedure, so it is in compliance with the law, however, sometimes crimes may arise in the process. Bankruptcy fraud happens when a firm purposely omits or fails to report information about its finances when applying for or during the process of bankruptcy. It is a white-collar crime (as it is financially motivated) which usually consists of the concealment of assets from the debtor to avoid liquidation during the proceedings. Not only that, some other examples may be: the filing of false information, conflicts of interests, and bribery, however, the crimes in this category are

dependent on the jurisdiction in question. In fact, filing multiple times in different countries does not constitute a crime itself but the applications may violate some provisions in some countries. An important distinction must be made with the concept of strategic bankruptcy, which occurs when a solvent company uses bankruptcy law to obtain business advantages other than the inability to pay debts. It must be highlighted that this is not a criminal act since it puts the firm in a real, ”legal” state of bankruptcy, however, it may still be detrimental to the company. During the bankruptcy procedure, it is the debtor's obligation to provide all the correct information, in full, and to cooperate with the authority in charge of the case. As a consequence, the debtor must disclose all the assets of the firm and it is for the creditors to decide whether a particular asset has value. Hiding assets or unpaid debts constitutes a form of bankruptcy fraud itself. The whole process is carefully monitored by the official receiver (the authority in charge of the case): from the assessment of the financial position of the firm before the bankruptcy to the completion of the process. Bankruptcy fraud is a serious crime and the fine may include an extension of the initial restriction period, a monetary fine, prosecution or even detention.

1.3 Differences in the bankruptcy legal frameworks around the world

As mentioned above, a company declares bankruptcy when it is not able to cover its debts or obligations. This will start a legal procedure that essentially aims at liquidating the firm's existing assets to repay the creditors. Generally, this is a worldwide definition, however, there are some differences in each country as the legal framework on this matter is not uniform. For example, in some countries (like the United Kingdom) bankruptcy is only applied to individuals, while firms undergo different insolvency procedures (such as liquidation). On the other hand, in other countries (like the United States) bankruptcy is applied more generally to formal insolvency proceedings. In the next sub-paragraphs, the legal framework for the United States, the United Kingdom, and Australia will be presented.

1.3.1 United States

In the United States, bankruptcy is ruled by federal law and all the provisions on the subject are contained in the U.S. Bankruptcy Code. There are various chapters dedicated, however corporate bankruptcy is treated in Chapters 7 and 11 of the code. The choice of the chapter applied to the case generally provides some clue as to whether the average investor will get back all, a portion, or none of their financial stake and the order in which the payment will be made, however, this will vary depending on the case.

When a company files for Chapter 7, the U.S. Securities and Exchange Commission states that ”the company stops all operations and goes completely out of business. A trustee is appointed to liquidate (sell) the company’s assets, and the money is used to pay off debt”. Clearly, debt is treated differently according to the risk exposure, with the less risky being paid first. This happens because there is a risk-return trade-off, for example, stockholders can gain a return when the share price of the firm increases while on the opposite, they can lose money when the share’s price falls. On the other hand, bondholders forego the opportunity of excess profits from the stock appreciation in order to gain a safer, regular, streak of specified interest payments on their bonds. This places stockholders after bondholders in a bankruptcy repayment scheme. Above all, secured creditors have absolute priority with respect to the other creditors. They accept very low interest rates in exchange for the added safety of corporate assets being pledged against corporate obligations.

On the other hand, Chapter 11 allows the firm to reorganize itself and does not put the company out of business, in the hope that its situation of insolvency will be fixed. This is the preferred choice for those corporations that need time to restructure unmanageable debt. This is the most complex and most expensive bankruptcy proceeding and it is therefore seen as the ”last resort” measure. Chapter 11 allows the firm to start fresh following a reorganization plan where all the necessary obligations must be fulfilled. The firm will be assigned to a committee operating in the best interests of the creditors and the stockholders and will be the authority responsible for the creation of the reorganization plan. In the Chapter 11 reorganization, the SEC states that ”bondholders will stop receiving interest and principal payments, and stockholders will stop receiving dividends. If you are a bondholder, you may receive new stock in exchange for your bonds, new bonds, or a combination of stock and bonds. If you are a stockholder, the trustee

may ask you to send back your old stock in exchange for new shares in the reorganized company. The new shares may be fewer in number and may be worth less than your old shares. The reorganization plan will spell out your rights as an investor, and what you can expect to receive, if anything, from the company."

Finally, Chapter 11 is definitely better than Chapter 7 from an investor's point of view, however, the reorganization process usually works just for a small number of firms, and other than being a long process, it is rare that the firm will return to the same profitability levels of its pre-bankruptcy state.

1.3.2 Australia

In Australia, two types of bankruptcy can be identified, depending on the subject who starts the filing. The first one is called voluntary bankruptcy and it happens when the debtors file a petition with the relevant court to go bankrupt. This is called a debtor's petition. The second one is called involuntary bankruptcy and it is demanded by the creditors of the firm who have been unable to recover their credits for a total of at least $ 5,000. The creditors will present a claim to the court which will declare the firm in the state of bankruptcy and will issue a sequestration order. In order to claim bankruptcy, a debtor must have at least $ 5,000 in debt and needs to file a petition to the Australian Financial Security Authority (AFSA), which is the competent national authority for bankruptcy applications. If compliant, the AFSA will appoint a trustee to oversee the debtor's finances. This process will produce effects on the firm, which will be considered legally bankrupt for three years from the day of declaration and will appear on a credit report for five years. During these three years, the debtor will have to comply with certain restrictions like the obligation to provide details of assets and income to the trustee. Similar to the US, secured debt and unsecured debt are treated differently. Finally, when the bankruptcy procedure has been completed, the firm will be released from almost all of its debts excluding some categories, like court fines.

1.3.3 United Kingdom

In the United Kingdom, bankruptcy only applies to individuals (including sole traders and members of partnerships) whereas companies that are unable to repay their debts are identified as insolvent and must undergo liquidation or administration. The differ-

ence between these two approaches is that company administration aims to help the firm to repay its debts in order to escape insolvency (if possible), while liquidation is the process of selling all the assets (in order to repay creditors) before dissolving the firm from the market. Obviously, a firm will always try to enter administration rather than liquidation since it will allow the company to continue running its business, however, an unsuccessful company administration could end in the liquidation of the company. The most important advantage of a company administration is that all legal actions against the company are suspended during the period of administration. When granted, an insolvency practitioner is appointed as administrator, taking full control of the company operations temporarily. For that period, the administrator will formulate and propose a recovery plan to the creditors of the firm acting in their best interests but still trying to help the firm repay as much debt as possible. Talking about company liquidation, when the firm is declared insolvent, directors are legally obliged to stop trading and must act in the creditors' best interests to avoid a further deterioration of the company's financial position. The type of liquidation closest to an American bankruptcy is a Creditors Voluntary Liquidation (CVL), which happens when the firm cannot repay its debts and a write-off of all the company's unaffordable unsecured debts is undertaken. This is a form of "compulsory liquidation" and it is a court-based procedure that aims at closing the company. This will allow the directors to free their position.

1.4 The advantages and disadvantages of filing for bankruptcy

Even though it represents a negative event for the firm, going bankrupt can be the right choice in certain situations. First of all, bankruptcy is a proceeding that releases the firm from debts that cannot be paid at that point in time while allowing creditors to benefit from the repayment of their credits through the liquidation of the firm's assets. In theory, this should lead to an improvement of the economy such that the debtor can enhance its creditworthiness and the creditor can finally collect (at least in part) his credits. In case of successful debt repayment, a discharge order can be granted to the debtor, which allows him to legally stop repaying his obligations as stated in the order and prevents creditors from any sort of collection activity. In addition, bankruptcy can

free the firm from old tax liabilities (older than 3 years). Last but not least, going bankrupt allows the firm to gain a fresh new start so that it can improve its credit score instead of trying to remain in business with a very hard time obtaining lines of credit.

On the other hand, while there are some benefits of going bankrupt, the negative consequences are quite important and this is why it is usually a last resort measure. The most important consequence is a large hit to the firm's credit score which will remain for 7-10 years on the credit report of the company (hence a long time for a firm's life). This will hinder the company's ability to get financing which is crucial to make investments and in general to continue the firm's operations. At the same time, this gives a bad reputation to the firm and a general sense of untrustworthiness since the name of the failed firm will appear publicly in court records. Other than that, even if old tax liabilities can be written off, it must be said that most tax debts cannot be discharged and in case the firm needs to file for bankruptcy again, it will need to wait a number of years before it is allowed.

All things considered, when debts are too large to manage, bankruptcy is definitely the right choice, since the alternatives would be a liquidation of all the assets and legal judgments for non-payment or breach of contract. Even though bankruptcy will damage the firm's reputation and creditworthiness, it is still a better legal procedure than the above-mentioned. On the other hand, when the firm's financial position can still be recovered, some alternative solutions (like debt restructuring) may be the better choice.

1.5 Debt restructuring as an alternative solution

Unlike in the past, modern insolvency legislation does not aim anymore at the elimination from the market of the insolvent entity but it aims more at modifying and improving the financial and organizational structure of the firm so that it can continue to operate in the market. This is why alternative solutions to bankruptcy have been preferred such as the debt restructuring of the firm. As the name suggests, debt restructuring is a process initiated by companies facing bankruptcy with the aim of restructuring the debt. It usually consists of a negotiation with the creditors of the firm with the aim of obtaining better lending agreements. This may be an extension of the liability's due date or an agreement to reduce the interest rates on loans. This is a very important process for the firm and it constitutes a much better solution to insolvency (with respect to bankruptcy)

for both the debtors and the creditors. On one hand, the debtors can increase their chances of repaying the liabilities and staying in business, hence avoiding bankruptcy (which would also come at a greater cost). On the other hand, creditors will receive more through a successful debt restructuring than through bankruptcy or liquidation so it is in their best interests to work with the firm in order to find an agreement. As stated above, there are multiple ways to perform a debt restructuring and some of these are debt-for-equity swaps, callable bonds, and ”haircuts”. First, debt-for-equity swaps happen when creditors agree to ”swap” a portion of the firm’s outstanding debt in exchange for equity (a stake in ownership). This is a preferred solution for big firms with a significant amount in both assets and debt, for which bankruptcy would be a huge loss. Second, the firm can issue callable bonds to protect itself when it is not able to cover interest payments. The peculiarity of these bonds is that they can be redeemed earlier than maturity when there are periods of decreasing interest rates. In this way, the firm performs a debt restructuring by taking advantage of the debt at lower interest rates. Lastly, the firm may negotiate with its bondholders to ”take a haircut”, meaning that a part of the balance or a portion of outstanding interest payments will not be repaid.

1.6 Relevant financial indicators

When a firm files for bankruptcy, it never happens ”out of the blue”, instead some clear signs of financial distress can be spotted prior to the filing. Financial statements include the business activities and the financial performance of the firm, so they are the obvious primary source for objective and relevant information about the company’s financial health. Anyway, they are not the only source because other signals can be found in other documents and events related to the firm. In the next paragraphs, some examples of bankruptcy warning signs will be presented.

1.6.1 The third-party auditor’s report

One example is the third-party auditor’s report which is a document published together with the quarterly and annual financial statements of the firm and it is written by third-party auditors who give an objective opinion on the firm’s performance. In the report, one warning signal can be a mention of discrepancies in the firm’s accounting practices

(like the method of accounting for costs) or the expression of uncertainty about the firm's future. Another sign can be a change in auditors, which is a common occurrence when the firm is in distress. This may be linked to a deterioration in the relationship between the auditor and the client company or to a general disagreement about the reliability of the company's accounting.

1.6.2 The managerial and operational areas of the firm

Another area of interest can be the managerial and operational framework of the firm. This is particularly important for closed companies, which do not disclose their financial statements to the public, and business information could potentially be the only source of warning signs about the company's health. One source could refer to changes in the market environment. These can be economic downturns, the entrance of a strong competitor in the market, or a shift of customer habits towards other brands. In general, this comprehends the environment in which the company operates: market trends, clients, competitors, suppliers, and so on. Another sign can be a change in the business strategy. It is very unusual and dubious when a firm shifts to a completely different business activity (like a production change to a completely different product) and for this reason, it should be taken into account. Another reason can be a sudden drop in the market prices of the products it produces. This is linked to the company's urge to increase sales and cash which may be needed to pay debts. An even worse sign is when the firm sells core business assets, which highlights the state of liquidity urgency. Linked to this, it should be noted when products lose their quality since the firm will try to cut costs to avoid bankruptcy and it will usually decrease the quality of the products first. Last but not least, when key executives leave the firm, it should be investigated.

1.6.3 Financial statements

Undoubtedly, the most important source of information about the firm's performance and activities can be found in financial statements. These documents are reported periodically (at least annually but many firms issue them even quarterly) so they are always up to date and reflect the firm's financial status in different areas:

- the Balance Sheet shows the financial position of the firm and displays the assets,

liabilities, and shareholders' equity;

- the Income Statement shows the company's income and expenditures for the period and it is particularly useful because it highlights if the firm has recorded a profit or a loss;
- the Cash Flow Statement displays all the cash inflows and outflows for the period;
- the Statement of Retained Earnings shows the cumulative business earnings after the payment of dividends. It also displays the change in the retained earnings account between the opening and closing periods on each balance sheet.

The main statement that needs to be checked is the Cash Flow Statement. When cash outflows exceed cash inflows for a prolonged period of time this means that the firm is running low on cash and it could be insufficient to cover the various obligations. If the company is not able to raise some capital from equity investors or lenders, it could potentially be in serious trouble. Especially cash flows from operations, when negative, are a signal that the firm cannot generate enough cash to sustain and further grow its operations and for this reason, it requires external financing. Linked to this, interest payments on loans can put pressure on cash flows, especially for distressed companies which will have to pay higher interests in order to compensate for their increased risk of default.

1.7 Financial ratios to spot bankruptcy

Financial ratios are probably the best indicators for assessing and comparing a firm on an objective scale. They are relative magnitudes of two selected numerical values derived from financial statements which are expressed either in decimal value or as a percentage. Financial ratios are divided into categories depending on the area they measure and these are:

- liquidity ratios: they measure the debtor's ability to pay current obligations without raising external funds;
- activity ratios: they measure how fast a firm converts non-cash assets into cash assets;

- debt ratios: they measure the firm's ability to repay long-term debt;
- profitability ratios: they evaluate the company's efficiency at generating profits and value for the shareholders;
- market ratios: they measure the cost of issuing stock and the investor's response to owning stock. They are focused on the return on investment for shareholders and on the relationship between return and value of an investment in a company's shares.

The most important feature of financial ratios is their comparability between competitors, industries or different time periods of a firm and can in fact be crucial when analyzing the financial position of a firm in distress. Among all, some are of particular importance when considering a distressed firm and these are: the current ratio, the cash flow to liability ratio, the liability to equity ratio, and the cash flow to sales ratio.

First of all, the current ratio takes the form:

$$Current\ Ratio = \frac{Current\ Assets}{Current\ Liabilities}$$

and is considered one of the main liquidity ratios to assess the business' solvency over the next year. Essentially, it measures the firm's ability to cover its liabilities that fall due within the next year. Generally, a current ratio of 2 or higher indicates good liquidity, whereas a ratio less than 1 is a warning sign.

Second of all, the cash flow to liability ratio is computed as follows:

$$Cash\ Flow\ to\ Liability\ Ratio = \frac{Cash\ Flow\ from\ Operations}{Total\ Debt}$$

Given the importance of cash flows, this ratio is considered the single best predictor of bankruptcy. This coverage ratio expresses the theoretical time period that it would take a company to pay all of its outstanding debt if all its cash flows were dedicated to debt payment. A ratio higher than 1 indicates a good ability to cover debt whereas a ratio lower than 1 warns about a possible bankruptcy in the coming years, should not the firm take action.

Third, the liability to equity ratio takes the form:

$$Liability\ to\ Equity\ Ratio = \frac{Total\ Debt}{Total\ Shareholders'\ Equity}$$

It is used to measure financial leverage and essentially aims at showing the company's ability to repay financing obligations and it displays the structure of the company's financing, whether it is more external (liabilities) or internal (equity). This is a crucial ratio for lenders when granting credit. If this ratio has a high value, it shows that the company relies a lot on external debt, which questions the firm's ability to repay its obligations. Generally, a ratio of 1 is optimal, with an equal proportion in debt and in equity, however, this is very industry-dependent and it is commonly accepted that a value of 2 (or higher) is considered unhealthy.

Finally, the Cash Flow to Sales is computed as follows:

$$Cash\ Flow\ to\ Sales\ Ratio = \frac{Cash\ Flow\ from\ Operations}{Net\ Sales}$$

It defines the company's ability to generate cash flows from its sales and it is considered good when sales and cash flow increments are in line. When this does not happen, the firm may be inefficient at managing costs or receivables. To sum up, these are some indicators that may spot early a bankruptcy case, however, the best prediction comes from looking at all the possible information about the firm's internal performance and its position in the market.

1.8 Worldwide bankruptcy trends

Usually, bankruptcy filings are tied to the status of the overall economy. Events like pandemics, wars, or financial crises tend to load the economy with a considerable amount of stress and this increases the number of bankruptcy filings. Figure 1.1 shows the annual U.S. bankruptcy filings in the 40-years-period from 1980 to 2020 including the applications for all chapters (both, those referred to individuals and companies). The figure shows an overall increasing tendency with peaks in certain years:

- 1997-2000 with the Dot-Com Bubble. This period is also known as the tech bubble because it was linked to the growth and increasing use of the Internet. In 5 years (beginning in 1995) the Nasdaq Composite stock market index rose by 400% and then when the bubble burst, it fell by 78% leading to the bankruptcy of many online shopping and communication companies.

- 2005 was the year with the most bankruptcies in the U.S. This is due to the passing of a new piece of legislation called Bankruptcy Abuse Prevention and Consumer

Protection Act which entered into force on October 15, 2005. This law restricted the possibility to file for Chapter 7 so that firms and individuals could not clear their debts and obtain a ”fresh start” that easily. Instead, the law made it more likely to apply under Chapter 13 which sets a repayment plan for up to 5 years. For this reason, the following year (2006) counted the lowest amount of bankruptcy filings in the U.S. up to that point. Anyway, 2005 counted a record-breaking 1.7 million Chapter 7 filings (46% more than 2004) and 8% less applications for Chapter 13.

- 2008-2010 was the three years period when the Global Financial Crisis really hit the economy, which was the largest financial crisis after the Great Depression (back in 1929). Essentially, aggressive lending to low-income home-buyers, excessive risk-taking by financial institutions, and the burst of the American housing bubble all contributed to the development of this crisis. Mortgage-backed securities tied to American real estate and derivatives linked to these securities both collapsed in value and caused severe distress to financial institutions. The most crucial bankruptcy (which coincided with the burst of the crisis) was the American bank Lehman Brothers, which had an impressive $691.1 billion in assets value. Of course, that was not the only remarkable bankruptcy, some others are Washington Mutual, General Motors, and CIT Group.

- Lastly, 2020 actually had the lowest amount of bankruptcy filings throughout the period, with approximately 500,000 filings across all chapters. However, business Chapter 11 filings continued to grow year after year and in 2020 they recorded a 29% increase with respect to the year before. According to the Epiq corporate restructuring managing director, “the peak in Chapter 11 filings for the second and third quarter is due to preexisting distressed companies coupled with the onset of a zero-revenue environment. The federal backstop proved a vital lifeline for the stabilization of corporations to protect the US economy. This federal intervention created record-breaking capital deployment fueled by investors chasing yield as companies attempt to ride out this storm”.

Undoubtedly, the two most important events in recent years are the Great Financial Crisis and the COVID-19 Pandemic, due to their repercussions on the economic system.

In the next two subparagraphs, the effects of these two catastrophes will be presented, with a focus on bankruptcy filings.

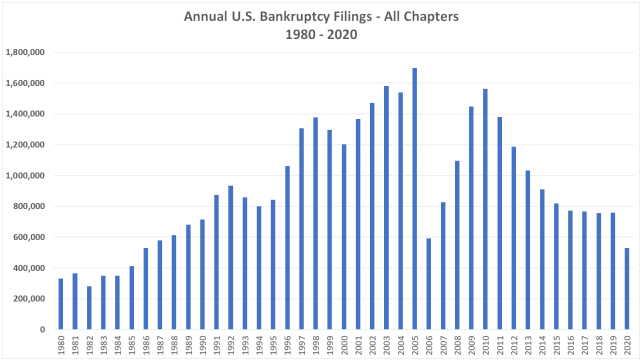

Figure 1.1 – Source: American Bankruptcy Institute

1.8.1 The Global Financial Crisis

The Global Financial Crisis began in 2007 and impacted the world financially during the coming years. Due to its impact, it is considered the most important financial crisis since the Great Depression. As aforementioned, the burst of the American housing bubble together with the increase of borrowers defaulting on their loans and deregulation in the financial sector, all brought to the development of the crisis. Unfortunately, the crisis was caused by a progressive deterioration of the whole financial system which could not really be predicted by any theoretical or empirical model except for minor signals. As a consequence, a sustained period of general market decline known as the Great Recession impacted the world, especially the most developed economies (like North America and Europe).

Talking about bankruptcies, filings increased by 74% over the two-year period ending June 30, 2009, and a total of 1,306,315 bankruptcy cases were filed in federal courts in 2009 with respect to 751,056 in 2007. Figure 1.2 shows the largest bankruptcies in the history of the United States (in terms of asset size) and it can clearly be seen that 50% of the companies in the graph went bankrupt during the Global Financial Crisis (from 2008 to 2011). This crisis caused the biggest number of large bankruptcies (companies with more than $100 million in assets) which amounted to 161 and 57 mega bankruptcies ($1 billion or more in assets). Not only that, the American investment bank Lehman Brothers is the biggest bankruptcy in history (with $691.06 billion in assets) and it was one of the major events of the Financial Crisis which caused a ”domino effect” on the global banking

sector and a 4.5% one-day-drop in the Dow Jones Industrial Average. In addition to that, 11 days after the collapse of Lehman Brothers, Washington Mutual declared bankruptcy with $327.91 billion in assets and became the U.S. largest bank failure. Finally, due to the devastating impact that the crisis caused on the financial sector, this brought regulators to enact new laws in order to preserve the financial system and prevent the occurrence of other banking crises in the future. In fact, in 2010 the American Congress enacted the Dodd-Frank Wall Street Reform and Consumer Protection Act which aimed at ”promoting the financial stability of the United States”, while the Basel Committee on Banking Supervision issued the Basel III accord which sets and increases international standards for bank capital adequacy, stress testing, and liquidity requirements.

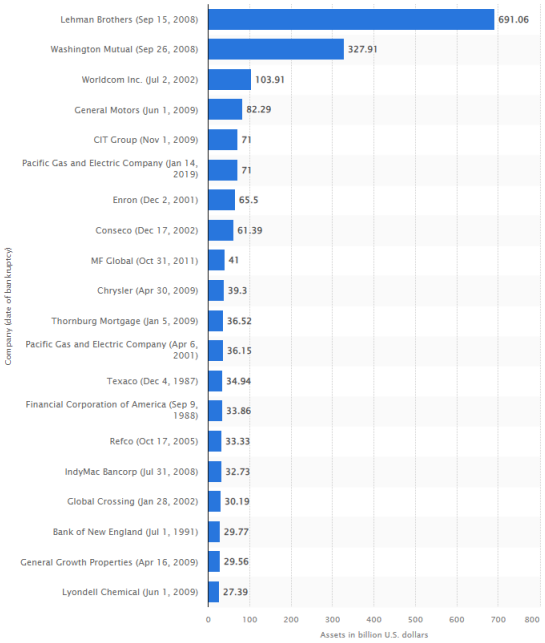

Figure 1.2 – Largest bankruptcies in the United States as of December 2021, by assets at the time of bankruptcy. Source: Statista.com

1.8.2 The COVID-19 Pandemic impact on large bankruptcies

The COVID-19 Pandemic is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) and it is considered the latest outbreak since the Spanish Flu (which happened approximately a century ago). The Pandemic caused a severe impact on the global social and economic system and it caused a global recession that is still ongoing. Essentially, the almost-global lockdowns that were imposed to restrict the circulation of the pandemic caused a supply chain disruption, which consequently brought widespread

supply shortages (including food shortages). As figure 1.3 shows, the COVID-19 Pandemic triggered a wave in large corporate bankruptcy filings in the U.S. which had not been seen since the Global Financial Crisis. The number of large bankruptcies in 2020 is second only to 2009 and filings by companies with $1 billion (or more) in assets were the highest since 2005, the year with the most bankruptcies in U.S. history. The figure also shows that, although remarkable, the bankruptcy wave driven by the pandemic was shorter than that caused by the Global Financial Crisis.

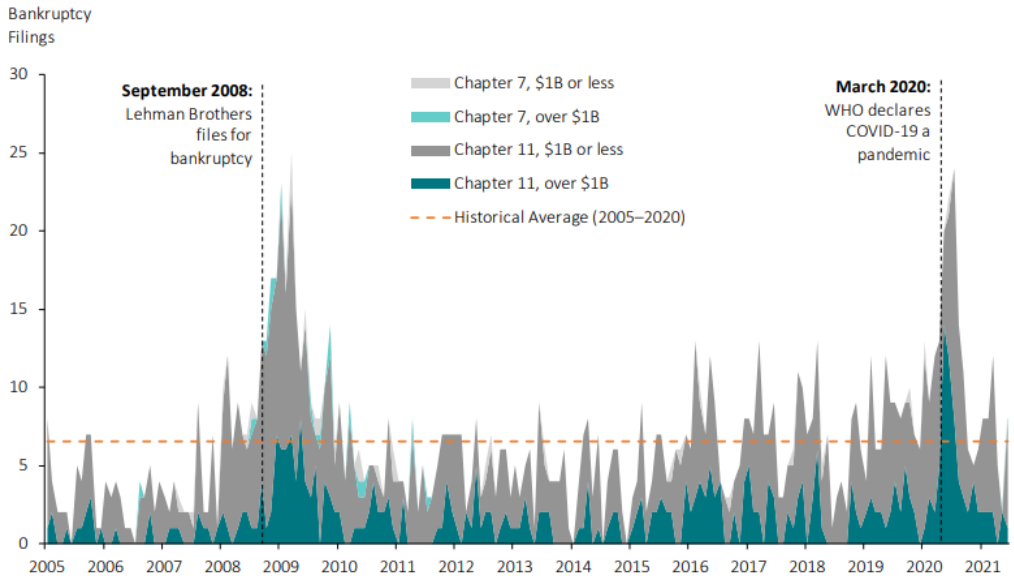

Figure 1.3 – Monthly Chapter 7 and Chapter 11 Filings 2005-1H 2021. Source: Cornerstone.com

Figure 1.4 demonstrates that as time passed and solutions were proposed both for the cure of the disease and for the global economic recovery, the overall number of filings decreased in the first half of 2021. In 2020, a total of 155 big bankruptcies (more than $100 million in assets) were filed, against 128 and 161 back in 2008 and 2009. There were 60 mega bankruptcies (more than $1 billion in assets), which is the highest number since 2005.

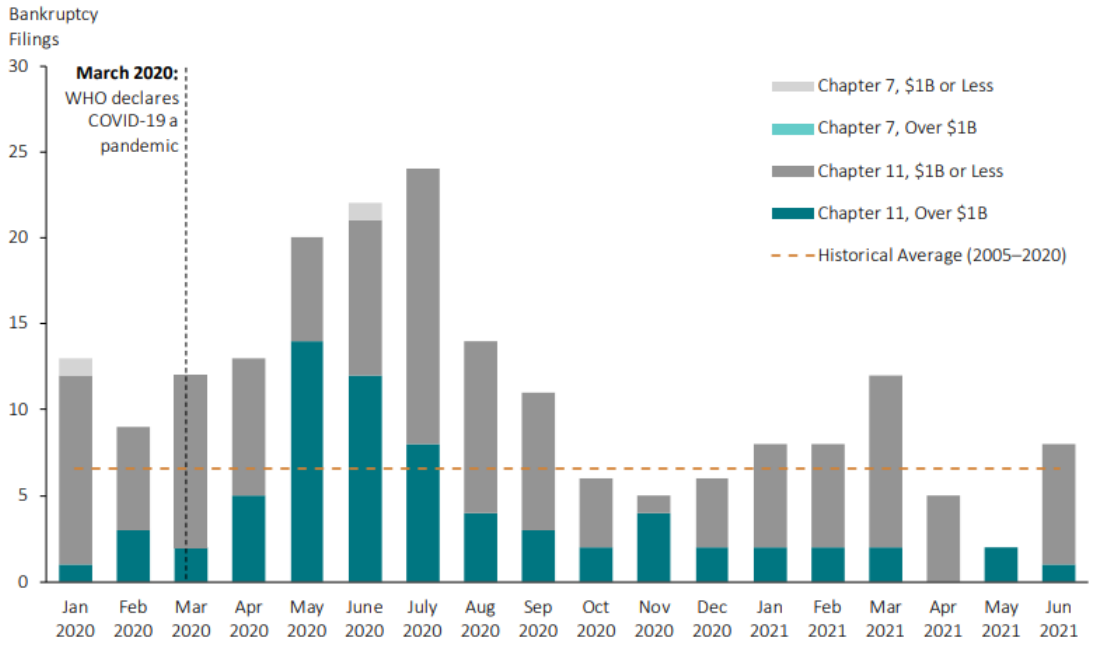

Figure 1.4 – Monthly Chapter 7 and Chapter 11 Filings 2020-1H 2021. Source: Cornerstone.com

The mining, oil, and gas industry was the one that suffered the most during the pandemic due to the already low oil prices which further collapsed and brought 44 new large bankruptcies in 2020. Another sector that was damaged is the retail trade, which counted 31 bankruptcies in 2020 mainly due to lockdowns and reduced demand for in-store shopping. These two industries combined for an incredible 48% of the largest bankruptcies in 2020, however, the circulation of the vaccine throughout 2021 decreased substantially these numbers and recovered (at least in part) the economy. The largest bankruptcy in this period was filed by The Hertz Corporation which admitted to having suffered the impact of the pandemic and declared bankruptcy in 2020, for a total of $25.84 billion in assets. However, after more than a year it was able to emerge from bankruptcy thanks to the equity capital injection of the new investor group of the company.

2 Bankruptcy prediction models

2.1 Introduction

A company is in financial distress when it cannot generate enough revenues to pay its obligations. According to Bruynseels & Willekens, there are different levels to this: when it is mild, the firm may face temporary negative cash flows and it may be insolvent or in default of some of its obligations. The highest stage of financial distress happens when the firm goes bankrupt, which leads to the discontinuity of the company's activities and creates repercussions for its stakeholders. This is why corporations need reliable business failure prediction models. Bankruptcy prediction is the art of predicting bankruptcy and various measures of financial distress of public firms. It is a large field of study which involves many aspects of a firm, such as the financial, managerial, and accounting areas. This is relevant for a number of subjects, mainly:

- the shareholders of the firm, since they own the firm's stock and have a financial interest, expecting a good performance. In particular, they pretend a share price increase so that they can sell their stack at a higher price;
- the creditors of the firm, since they are particularly interested in the creditworthiness of the company and expect repayment from the line of credit that they granted. Especially when a firm declares bankruptcy, it is not able to repay its debts and has low creditworthiness;
- the investors of the firm, whether they invest in stock or equity, they are interested in a good performance because they want higher returns from their investments. Like shareholders, when owning stock, they expect a share price increase. When owning bonds the same concept applies: good performance indicates a good credit rating which reflects the firm's ability to repay the debt. When a firm goes bankrupt, it probably defaulted on its bonds and that is why a bad credit rating is applied (usually a D, which stands for default).

Usually, better predictions come with more data: in fact, public firms are the best type of companies to be predicted since they provide a lot of accounting ratios and other ex-

planatory variables. As a consequence, bankruptcy prediction is appropriate for testing increasingly sophisticated, data-intensive forecasting approaches.

2.2 The classic bankruptcy prediction models

In the past, studying bankruptcy prediction involved using statistical tools to identify the effects of certain variables (usually accounting ratios) on the firm's probability of going bankrupt. This field of study dates back to 1932 when FitzPatrick analyzed the financial ratios of 20 pairs of firms (one surviving and one failing) and interpreted trends in the ratios. In 1966, Beaver applied t-tests to evaluate the importance of individual accounting ratios within a similar pair-matched sample. Among 30 ratios he identified that 3 are significant for predicting bankruptcy (namely: Total Assets to Total Debt, Net Income to Total Assets, and Cash Flow to Total Debt). In 1968, Altman developed one of the most important early models to predict bankruptcy (still used nowadays), which consists of a formula made of 5 financial ratios that would assign a score to the firm, which depending on the value, would be considered in risk of bankruptcy or not. In 1980, Ohlson applied logistic regression to a large sample that did not involve pair-matching. Subsequently, in 1984 Zmijewski applied the probit function to the sample and used three ratios to build his model. Nowadays, according to a literature review made by Jackson and Wood in 2013 analyzing 15 popular models in this field, there is a vast range of possibilities: from the univariate models of Beaver to those multidimensional developed by Altman and Ohlson to more modern models based on market data (such as an option valuation approach) or using machine learning methods in order to obtain the most accurate predictions. In the next paragraphs, the most famous bankruptcy prediction models will be presented and critiqued.

2.2.1 Altman Z-Score

The Altman Z-Score is a bankruptcy prediction model which was released by Edward Altman in 1968 and is considered the first important model on this matter, which still finds application to this day. The model was discovered to assist investors in defining how close the firm is to bankruptcy, which was considered a somewhat confusing and time-consuming process up to that moment. As time passed, the model has proved to

be more suited to provide information about the overall financial health of a company instead of being used solely to predict bankruptcy probability as better methods have been developed. In 2012, Altman presented an updated version of the model (the Altman Z-Score Plus), which is much more comprehensive: in fact, it can be used for both public and private companies, manufacturing and non-manufacturing companies, U.S. and non-U.S. companies and it is useful also to evaluate corporate credit risk.

Essentially, the model is a variation of the traditional z-score in statistics (a numerical measurement of a value's relationship to the mean in a group of values) and aims at predicting the probability that a firm will go bankrupt within two years. The formula of the model is as follows:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

$where$:

$X_1 = \frac{Working\ Capital}{Total\ Assets}$

$X_2 = \frac{Retained\ Earnings}{Total\ Assets}$

$X_3 = \frac{Earnings\ Before\ Interest\ and\ Tax\ (EBIT)}{Total\ Assets}$

$X_4 = \frac{Market\ Value\ of\ Equity}{Total\ Liabilities}$

$X_5 = \frac{Sales}{Total\ Assets}$

It uses five financial ratios that can be computed from data found in the firm's annual financial report, which provide a good overview of the firm's status since they cover profitability, leverage, liquidity, solvency, and activity of the company. X_1 compares the net liquid assets to the total assets of the firm. Working capital is defined as the difference between current assets and current liabilities and the ratio provides information about the company's short-term solvency. X_2 shows the proportion of total assets funded by the accumulated earnings of the period in question (usually one year). The ratio indicates the management's tendency to use retained earnings to reinvest in the firm instead of paying dividends. X_3 shows the company's level of efficiency in generating earnings with its assets. X_4 is a solvency metric and shows the extent to which the firm's assets can decline in value (measured by the market value of equity plus debt) before the liabilities exceed the assets and the firm becomes insolvent. X_5 gives information about the company's ability to generate sales using its assets.

As shown in figure 2.1, depending on the value of the z-score, the firm can find itself in one of three zones:

- when the score is below 1.8 the firm is in the ”Distress Zone”, which means that is probably going to fail in two years. If an investor owns stock of this firm, it is suggested to him to sell it before the company goes bankrupt and he could potentially lose the whole investment amount. Altman later specified that when the z-score is close to 0, investors should really worry about the future of the firm.

- when the firm finds itself with a score ranging from 1.8 to 3.0, this means that it is in the ”Grey Zone” and it has a moderate probability of going bankrupt. This does not give so obvious information about the firm's stock, however, when the value is closer to 1.8 the investor should consider selling it, and vice versa when it is closer to 3.0 he should consider buying.

- when the score is above 3.0 the firm is in the ”Safe Zone”, which means that bankruptcy is unlikely in this case. Investors should consider buying the company's stock since the firm is unlikely to go bankrupt in the two coming years.

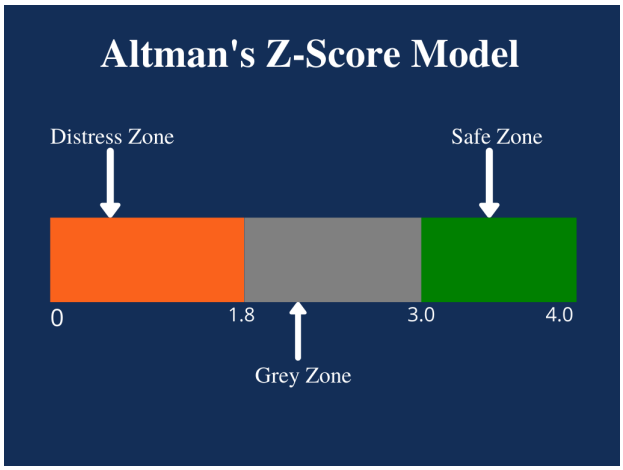

Figure 2.1 – Z-score ranges of financial stability. Source: CorporateFinanceInstitute.com

When the model was first tested in 1968, it recorded a 72% accuracy in predicting bankruptcy two years earlier, with a 6% of Type II error (false negatives). In future tests performed in three periods in the following 31 years, the model recorded 80%-90% accuracy prediction one year before the event with a Type II error of 15%-20% (Altman, 2000).

In the past, the Altman Z-score has been heavily critiqued, since it represents one of the pioneering models of bankruptcy prediction. In 1970 C.G. Johnson defined it as “largely descriptive statements devoid of predictive content (...) Altman demonstrates that failed and non-failed firms have dissimilar ratios, not that ratios have predictive

power. But the crucial problem is to make an inference in the reverse direction, i.e., from ratios to failures". According to Eidleman, from 1985 onwards the z-score has been successfully implemented as a tool for loan evaluation and gained popularity in the market. This changes the scope of application of the model, from a bankruptcy prediction tool to a "credit risk assessment method". The z-score can appear misleading also when one-time write-offs happen and its value can change considerably from quarter to quarter, suggesting that one company is on the verge of bankruptcy when it really is not. Another weakness of the score is that it is not immune to the misrepresentation of financial data: in fact, it is very common for firms in financial distress to manipulate accounting data to display a better financial situation than what it really is. In addition, the model is not very useful for new firms with poor earnings (like start-ups), as they will obtain a low score in any case. Another big fallacy is that the z-score does not take into account cash flows directly but only gives a hint with the use of the working capital in the first index. Talking about the Global Financial Crisis, back in 2007 (just before the burst of the crisis) Altman stated that the corporations' risks were increasing considerably. The American professor believed that the crisis would develop due to high rates of corporate defaults, which although it was not the primary mover (the crisis burst due to mortgage-backed securities), led to corporate defaults in 2009 at the second-highest rate in history. Table 2.1 shows the z-scores for 5 years for some famous defaults that happened during the 2008 Financial Crisis, together with the rating at the time of default given by the most famous credit rating agencies. As the table shows, in all 5 years for all 6 firms, the z-score was below 1.8 and often negative or close to 0. This means that the Altman Z-Score did a good job at recognizing these firms as in distress, however, this cannot be said for the credit rating agencies which gave mostly good ratings for the firms at the time of default (only Ford Motors and MF Global showed signs of "instability"). This is especially true for Lehman Brothers, which is considered the biggest bankruptcy in history and resulted in the main event which gave the start to the Global Financial Crisis.

Defaulter	Amount of Liabilities	Date of Default	Z-Score			Rating at the time of default			The Consequences
	(in $ billion)		In year of default	1 year prior	2 years prior	S&P	Moody's	Fitch	
Bear Sterns	387	31 July 2007	0.29	-0.79	0.45	AA a A	A1 a A2		Acquired by JP Morgan Chase
AIG	807	16 Sep 2008	-1.03	-0.07	-0.02	AA- a A-	A1 a A2	AA- a A-	Bailed out by US Government
Lehman Brothers	392	23 Sep 2008	0.06	0.09	0.03	AA, A1	P1 & A1	AA– & F1+	Bankrupt
Washington Mutual Bank	303	25 Sep 2008	-0.35	-0.3	-0.07	A- & A2	Baa1 & P2	A- & F2	Acquired by JP Morgan Chase
Ford Motors	132	6 Apr 2009	1.32	1.03	1.23	CC	Caa1, B3	CCC, BB	Revived
MF Global	51	31 Oct 2011	0.23	0.47	0.37		Baa2 a Caa	BBB a BB+	Bankrupt

Table 2.1 – Altman's prediction on famous defaults of the Global Financial Crisis. Source: Wikipedia.com

2.2.2 Ohlson O-score

The Ohlson O-score is a multi-factor financial formula developed by the American professor James Ohlson in 1980, which aims at predicting bankruptcy as an alternative to the Altman Z-score. The model is the result of a 9-factor linear combination of coefficient-weighted business ratios which can be found in the corporation's financial statements.

The model has the formula:

$$T = -1.32 - 0.407log\frac{TA_t}{GNP} + 6.03\frac{TL_t}{TA_t} - 1.43\frac{WC_t}{TA_t} + 0.0757\frac{CL_t}{CA_t} - 1.72X - 2.37\frac{NI_t}{TA_t} - 1.83\frac{FFO_t}{TL_t} + 0.285Y - 0.521\frac{NI_t - NI_{t-1}}{|NI_t| + |NI_{t-1}|}$$

$where:$

$TA = Total\ Assets$

$GNP = Gross\ National\ Product\ price\ index\ level\ (in\ USD,\ 1968 = 100)$

$TL = Total\ Liabilities$

$WC = Working\ Capital$

$CL = Current\ Liabilities$

$CA = Current\ Assets$

$X = 1\ if\ TL > TA,\ 0\ otherwise$

$NI = Net\ Income$

$FFO = Funds\ From\ Operations\ (pre - tax\ income + depreciation)$

$Y = 1\ if\ the\ firm\ recorded\ a\ net\ loss\ for\ the\ last\ two\ years,\ 0\ otherwise$

The model can also be used to identify the probability of the firm's default, taking the form:

$$p(default) = \frac{e^{O-score}}{1+e^{O-score}}$$

Each factor of the equation measures a different area of the firm and two of these factors are widely considered to be dummy variables since their value is usually 0. The first factor measures the "adjusted size" of the firm, taking the logarithm of the company's

total assets adjusted for inflation (measured by the GNP price-level index). According to this metric, smaller companies are thought to be more prone to bankruptcy than others. The second factor is a leverage measure and it is useful to identify the company's level of debt: the higher it is, the more the firm is at risk of economic shocks. The third measures working capital, assuring that a firm should have enough liquidity to cover short-term debt and upcoming operational expenses to avoid bankruptcy. The fourth basically has the form of an "inverse current ratio" and is another measure of liquidity. The fifth is a "discontinuity correction for the leverage measure" which has the form of a dummy variable taking value 1 when total liabilities exceed total assets, and 0 otherwise. However, this is an unusual case to account for extreme leverage positions, that is why this value is often 0. The sixth factor measures the return on assets which determines how effective the company is at generating profit using its assets (it is assumed to be negative for bankrupt companies). The seventh element is the Funds to Debt Ratio which assesses the company's ability to finance its total liabilities only using operational income. If the ratio is lower than 1, this is a warning sign. The eighth factor constitutes the other dummy variable of the model and represents a "discontinuity correction for return on assets": it takes value 1 if income was negative in the last two years, otherwise it is zero. Finally, the last factor displays the change in net income and is used to account for any potential continuous losses over the most recent spans of the company's history. Similarly to the z-score, the o-score measures the company's probability of bankruptcy within two years, however, the ranges are different: when a company scores more than 0.5, it is projected to go bankrupt within two years.

Originally, the Ohlson O-Score was tested on a sample of about 2000 companies, whereas the Altman Z-Score just used 66. Due to its testing on a much larger sample, the o-score is considered a better bankruptcy predictor in a 2-year period than the z-score. As mentioned above, the original z-score obtained a 70% accuracy, which increased up to 90% in later tests. The o-score obtained as high as 96% when tested, however, no model has 100% accuracy and there are outside factors that simply cannot be taken into account. At the same time, modern models like the "hazard-based model" proposed by Campbell, Hilscher, and Szilagyi in 2011 have shown to be even more accurate. According to a 2001 study critiquing classic bankruptcy prediction models in current research[10], problems may arise when these types of models are inappropriately applied. The study analyzed

the Zmijewski (1984) and Ohlson (1980) approaches using industries, time periods, and situations of financial distress that were different from those used when developing the models. The study proved that the accuracy of the models declined when using different time periods. Ohlson's model proved to be sensitive to different industries, whereas Zmijewski's did not. Additionally, both models proved to be insensitive to different financial distress situations. This suggests that these types of models should be used in the market scenarios that they were developed for, otherwise their accuracy will decrease substantially. At the same time, predicting bankruptcy is a "limiting factor" of these models and they proved to be better suited to evaluate various forms of financial distress.

2.2.3 Zmijewski's Score

The Zmijewski's Score is another well-known model used for bankruptcy prediction which was developed by Zmijewski in 1984. Together with the other two (Altman Z-Score and Ohlson O-Score), it falls under the category of the early models used to predict bankruptcy. Unlike the other two, Zmijewski had a more "parsimonious" approach in choosing the factors of the model and decided to use a probit model, a classification model based on the probability of the event. The formula was developed considering all the American listed companies in the period 1972-1978 (excluding financial companies) and has the form:

$$Zmijewski's\ Score = -4.336 - 4.513X_1 + 5.679X_2 + 0.004X_3$$

$$where:$$

$$X_1 = \frac{Net\ Income}{Total\ Assets}$$

$$X_2 = \frac{Total\ Liabilities}{Total\ Assets}$$

$$X_3 = \frac{Current\ Assets}{Current\ Liabilities}$$

Each one of the three variables was chosen to measure a different aspect of the firm. X_1 is commonly referred to as the ROA (Return On Assets), which is a measure of profitability and shows how efficient the company is at generating profit using its assets. X_2 is a leverage measure and defines the total amount of debt relative to assets owned by a company. While X_3 is a liquidity metric to check if a company is able to cover its short-term liabilities (due within a year). When the model was first developed, Zmijewski tested it on a sample of 40 failed and 800 stable firms. The higher the score, the higher the probability of bankruptcy for the company, and generally, when the score is

above 0.50, the firm is considered in danger. When tested, the model scored 99% accuracy, however - as mentioned above - when changing the market scenario in which the model was developed, its accuracy tends to decrease substantially. Finally, the model was critiqued for the low number of variables considered and for their collinearity (they were shown to be strongly correlated).

2.2.4 Comparing the three models

These three classical models offer a practical prediction of the risk of bankruptcy of the firm, using accounting values that can be obtained by the financial statements of the firms. All three offer ”decent” accuracy, however, among the three, the Ohlson O-Score seems the best option to predict bankruptcy for two reasons. First of all, the o-score is a more ”comprehensive” metric, as it considers nine factors with respect to five of the Altman Z-Score and three of the Zmijewski's Score. These factors include also inflation-adjusted total assets, short-term and long-term liquidity, profit before and after taxes for the current year and also previous years, and market value. The other two models are definitely simpler and can offer a prediction with the need for far less data, however, this hurts their accuracy. Second, the Ohlson O-Score historically scored higher when tested in comparison with the Altman Z-Score: the latter obtained 90% accuracy in later tests, while the o-score recorded as high as 96%. On the other hand, the Zmijewski model was usually neglected due to its excessive simplicity and collinearity among predictors. To sum up, these models can offer a practical view of the financial status of a firm and can offer a relatively accurate prediction of their risk of bankruptcy, using accounting-based data. However, their simplicity is also the biggest weakness of these models, which simply cannot be considered reliable primary methods of bankruptcy prediction as they are too restrictive on the data used and do not work well in market scenarios different from those in which they were developed. However, they are still considered valid secondary tools to confirm the results of more powerful models and that is why they are widely used still to this day.

2.3 Machine learning: the superior choice

The three famous models analyzed in the previous paragraphs have been among the first attempts of predicting bankruptcy, however, due to their simplicity, they cannot be used with much versatility. On the other hand, starting from the 1990s, supervised machine learning models have been widely used in classification problems due to their high predictive power.

A machine learning model is a computer program that has the task of finding patterns or making decisions based on an unseen dataset, with specific rules and data structures. The model performs the tasks by training it on a large dataset. During the training phase, the machine learning algorithm is optimized to find certain patterns or outputs from the dataset, depending on the specific task attributed to the model (model training). The output of the process is what constitutes the machine learning model. As mentioned, the model uses an algorithm to perform the tasks and such an algorithm is a mathematical method drawn from statistics, calculus, or linear algebra (such as logistic regression, random forest, or linear regression). Given its high versatility, machine learning approaches are applied to many fields of study: from medicine to email filtering or speech recognition, they are very useful when it is hard or simply impossible to develop conventional algorithms to solve certain problems.

Bankruptcy prediction is another field where machine learning algorithms are very useful since previous models like the Altman Z-Score, the Ohlson O-Score, and the Zmijewski's Score have been shown to be too simplistic which results in just a "moderate" predictive power and also not versatile at all since their reliability decreased dramatically when applied to different market context than those in which they were developed. Especially in subject matters like bankruptcy prediction where an important problem is being evaluated (in light of recent crises like the Global Financial Crisis where the failure of many financial firms was totally unexpected) and the bankruptcy of an important firm would result in a huge loss of money and a remarkable impact on the economy (like in the case of Lehman Brothers), models with high predictive power and versatility are much needed.

In addition, there has been an increase in the use of big data recently, which consists of the use of very large and/or complex datasets which simply cannot be analyzed with traditional data-processing softwares. In fact, data with high complexity (such as when

there are many attributes or columns) may lead to a high false discovery rate. There are many machine learning algorithms that can handle huge loads of data and perform very well in the field.

The task of predicting bankruptcy is a problem of binary classification, which aims at identifying which of a set of categories (bankrupt or non-bankrupt) an observation belongs. The algorithm that implements classification is called a classifier, which is the mathematical function implemented by a classification algorithm that sorts input data into a category. The problem falls under the category of supervised machine learning: an input dataset is provided to the algorithm and it is optimized to reach a set of specific outputs. Researchers are exploring which machine learning tools can accurately perform this classification (Wilson and Sharda (1994); Tsai (2008); Chen et al. (2011)): many academics combine statistical methods with machine learning approaches to obtain the best possible predictions. For example, Cho et al. (2010) developed a hybrid model choosing variables selected by decision tree and case-based reasoning using the Mahalanobis distance with weights. Chen et al. (2009) instead proposed a hybrid model which combined fuzzy logic and neural network. The final results demonstrate that the hybrid model has higher accuracy than the logic model. Another category of successful models on this matter is constituted by Support Vector Machines (SVM), in fact, Cortes and Vapnik (1995) create functions similar to discriminant analysis and provide a final successful prediction of corporate bankruptcy. SVMs are widely used by researchers to predict bankruptcy (Shin et al. (2005); Chaudhuri and De (2011); Sun and Li (2012)) and some scholars attempted at changing the algorithm to further improve the accuracy of the model: Chaudhuri and De (2011) used fuzzy SVMs to solve the problem and highlighted their efficiency, while Zhou et al. (2009) suggested a method to optimize parameters in SVMs. Artificial Neural Networks (ANNs) are another option and they have a similar structure to the usual neural network. In the ANN the input layer is the input variable and the output layer determines the output variable, with hidden layers between the first and the last. ANNs can be very useful in order to analyze non-linear relationships with respect to traditional models. The peculiarity of the ANN is that it can improve the accuracy of the model by changing the setting of the parameters.

2.4 The purpose of this study

2.4.1 Introduction

As analyzed in the previous sections, the event of bankruptcy is very important for the life of a company and for the economy as a whole. When a firm goes bankrupt, it hurts a wide range of subjects: the firm ceases to exist, the management of the firm loses money and the same happens for creditors and investors. In addition, recent events like the Global Financial Crisis or the recession caused by the COVID-19 pandemic brought even more attention to the problem as bankruptcy rates increased in those periods. This is why the problem of bankruptcy prediction is crucial nowadays: firms need reliable methods to assess in advance when it is at risk of bankruptcy. This can ultimately help the management team to spot inefficiencies inside the company so that they can fix them and avoid the worst-case scenario. In this thesis, I promote the use of popular statistical models in conjuction with machine learning techniques to predict bankruptcy, based on a dataset of Taiwanese companies. In the previous section, the relevance of machine learning methods in this field was highlighted, not only because they are very versatile but also because they are well suited for big data analysis and offer much better accuracy than classic models (like those of Altman, Ohlson and Zmijewski).

2.4.2 Presentation of the problem

The purpose of this study is to use famous statistical models and machine learning techniques to obtain the most accurate prediction of a firm's probability of going bankrupt. From a statistical point of view, this is a problem of binary classification in which observations (in this case companies) must fit one of two sub-populations (bankrupt or non-bankrupt). For this specific type of problem, the data is usually imbalanced since bankruptcy is a rather rare event and the vast majority of firms (of course considering a time period of relatively normal market conditions) are considered financially stable. This means that the dataset requires resampling since learning algorithms tend to perform poorly with imbalanced data and this would result exclusively in the prediction of the majority class, which is not what we want since we are more interested in detecting bankruptcies rather than stable companies, which constitute the vast majority of the dataset. As a result, the analysis gives particular care to false negatives (type II

error): the number of real bankrupt firms predicted as stable must be low. Resampling is done through various techniques, however for the purpose of this study we will consider random oversampling, random undersampling, and SMOTE (more on these in the next chapter). Lastly, since this is a classification problem we will use supervised learning methods in which the algorithm is provided with an input dataset, and is rewarded or optimized to meet a set of specific outputs. The classifiers will be logistic regression, LASSO regression, and random forest. Each and every model will be optimized (for example regarding the choice of the parameters) and will be evaluated with classification metrics like the confusion matrix, ROC-AUC curve, F-1 score, and so on.

2.4.3 The procedure of the analysis

The analysis is performed on a dataset of Taiwanese companies, including 6819 observations and 96 attributes. The first step consists of the pre-processing of the data which includes the upload to the software of the dataset and some basic information gathering about the data, such as the number of null entries or the exact number of firms in each category. The second step is mainly dataset rebalancing, applying the random oversampling, random undersampling, and SMOTE techniques to balance the two classes in order to make the learning algorithm work. The third step involves the selection of the model among logistic regression, LASSO regression, and random forest. The fourth step is the split of the data between a training set and a test set. Subsequently, the model is trained using the training set (which constitutes the majority of the dataset) to learn the functioning of the algorithm. The fifth step consists in testing the model on the testing set. In the sixth step, the model is evaluated using the traditional classification metrics: confusion matrix, ROC curve, and F1 score. Some others are used such as specificity, sensitivity, or accuracy however, they cover a ”secondary” role. Finally, the last step consists of the comparison of the results (using the abovementioned metrics) and the selection of the best model. Figure 2.2 illustrates graphically this process.

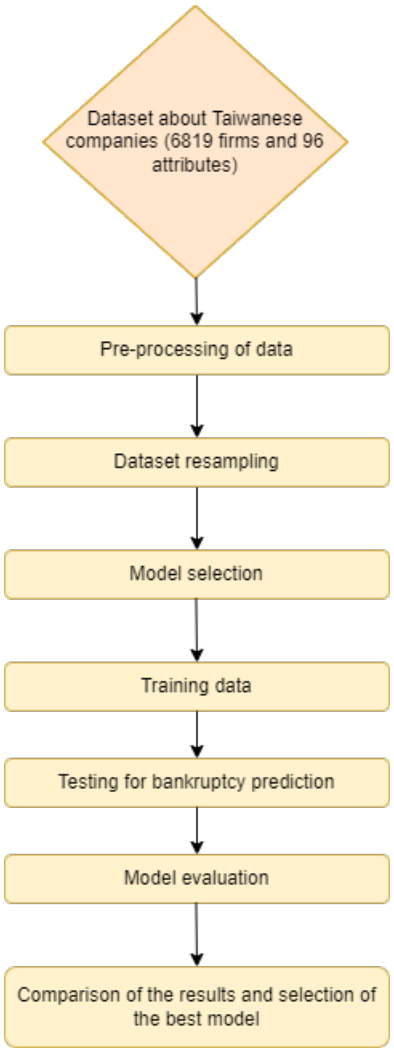

Figure 2.2 – Procedure of the analysis

2.4.4 The dataset used

The analysis is performed on a dataset collected by the Taiwan Economic Journal[1] from 1999 to 2009, which was obtained from UCI Machine Learning Repository[2][3] and company bankruptcy was defined according to the business regulations of the Taiwan Stock Exchange[4]. There is only one relevant paper using this dataset that performs a comprehensive study on financial ratios and corporate governance indicators on bankruptcy

[1] https://www.tej.com.tw/

[2] https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

[3] Deron Liang and Chih-Fong Tsai, deronliang '@' gmail.com; cftsai '@' mgt.ncu.edu.tw, National Central University, Taiwan

[4] https://twse-regulation.twse.com.tw/ENG/EN/law/DAT0201.aspx?FLCODE=FL007304

prediction[13].

The time period chosen is interesting since it can also partially take into account the effects of the Dot-Com Bubble and the Global Financial Crisis on the Taiwanese economy. The dataset presents 6819 observations and is multivariate, in fact, it comprehends 96 attributes: 95 predictors and 1 dependent variable. The predictors are basically 95 financial metrics and provide a good insight into the firm's financial health, measuring for example profitability, liquidity, operational efficiency, solvency, and so on. On the other hand, the first attribute is the dependent variable y which constitutes the class label and categorizes the status of firms. All these characteristics of the dataset make it optimal to perform classification on it.

3 The theoretical background of the analysis

3.1 Introduction

In this chapter, the theoretical background of the analysis is presented. The first few paragraphs cover the identification of the problem, starting with a general categorization and then going deeper. In the beginning, the differences between supervised and unsupervised learning are compared and we can see that bankruptcy prediction falls into the supervised learning category. More specifically, the problem belongs to binary classification and presents a severe imbalance in the class distribution. In order to fix this imbalance (which can hurt the model's accuracy), some solutions are proposed (mainly the use of resampling techniques but also ensemble methods). In the following section, the statistical models used in the analysis and their relative functioning are described. The chapter concludes by talking about the performance metrics used to evaluate these algorithms and discusses their pertinence considering the problem in question.

3.2 Supervised and Unsupervised statistical learning

Most statistical learning problems are subdivided into two categories: supervised and unsupervised. On one hand, in supervised learning for each observation of the variable measurement(s) $x_i, i = 1, ..., n$ there is a linked response y_i. Supervised learning models find a relation between the predictors and the response, in order to predict accurately the response for future observations (a predictive model) or to analyze the relationship between the predictors and the response variable (inference model). Linear regression and logistic regression are two examples of supervised learning methods. On the other hand, in unsupervised learning there is a vector of measurements x_i for every observation $i = 1, ..., n$, however in this case there is no association to a response y_i. This is more troublesome since there is no response variable which can supervise the analysis and for this reason, supervised methods like linear regression are not effective since there is no response variable to predict. Bankruptcy prediction is a type of problem which falls

under the category of supervised statistical learning, in fact, firms' financial ratios are linked to the response variable, and models in this field aim at using such ratios to obtain accurate predictions of the response variable, in order to assess the financial health of companies.

3.3 Overview of the problem

3.3.1 Binary classification and feature selection

Bankruptcy prediction is the art of predicting a firm's bankruptcy in advance with a certain accuracy so that the company can attempt to fix its weaknesses and avoid the event. As already mentioned, this type of analysis is performed mainly with the use of machine learning techniques based on statistical models, since the literature shows their effectiveness to solve this type of problem (Li, Y. and Wang, Y. (2018)). More precisely, from a statistical point of view, the problem falls under the category of binary classification. In statistics, classification is the problem of fitting an observation (or multiple observations) into a set of categories (sub-populations). The response variable is qualitative and takes values in one of the K classes. This is a common problem present also in other fields like medicine, when, given a set of symptoms, the patient is categorized as "ill" or "healthy". Usually, the observations are analyzed into a set of quantifiable properties, defined as explanatory variables or features: these are independent variables, the variation of which, affects the dependent variable. Instead, the dependent variable "contains" the categories to which each observation belongs, and these can be two (binary classification) or more (multiclass classification). Binary classification is often a better-understood task and is more common in real-life examples while multiclass classification is more complex and requires the combined use of multiple binary classifiers. Classification is implemented through a classifier: an algorithm that automatically orders data into one or more sets of classes. The classifier is built with a set of training observations $(x_1, y_1), ..., (x_n, y_n)$ and to declare its effectiveness, it needs to perform well both on the training data and on the test set (which was not used to train the model).

In particular, in bankruptcy prediction, the dependent variable is a dummy that can take only two values: 0 if the firm is stable and 1 if it is bankrupt. On the other hand, the features usually consist of financial ratios about the internal performance of the firm

(expressed as real numbers). Clearly, a high number of ratios is preferred, in order to have a complete overview of the firm's financial status, accounting for all the major operational areas. However, it often happens that some variables are not associated with the response and including them in the model increases its complexity uselessly. A procedure called feature selection removes these unnecessary variables from the model by setting the correspondent coefficient estimates to zero. This technique is performed for a number of reasons:

- simplification of models in order to make them more easily interpretable;
- shorter training times (a considerable amount of predictors could be very demanding from a computational aspect);
- in order to avoid the curse of dimensionality: the presence of noise features that are not associated with the response worsens the performance of the fitted model and increases the test set error as a result;
- to improve data's compatibility with a learning model class;
- to encode inherent symmetries present in the input space.

The main objective of feature selection is to identify redundant or irrelevant features and remove them from the model without much information loss. In order to do this, it is important to check the correlation among features (collinearity) to identify these variables.

3.3.2 Imbalanced classification

The first challenge when conducting bankruptcy prediction is recognizing that the problem falls under the category of imbalanced classification: too many firms are allocated to the "stable" class and too few are those bankrupt. This is surely linked to the event itself, in fact, bankruptcy is a rather rare event and is usually avoided and other solutions are preferred (when possible). In a problem of imbalanced classification, the distribution of observations across the known classes is unequal, and depending on its severity, it can seriously affect the predictive power of the model. As a matter of fact, most algorithms were designed with the assumption of equal classes' size and an imbalanced dataset hurts the accuracy of the model, especially for the minority class. This is troublesome since

the minority class is typically more important (because it expresses the occurrence of a rare event) and therefore the problem is more sensitive to classification errors for the minority class than those of the majority class. Technically, any dataset with an unequal class distribution is imbalanced, however, it is considered a problem when there is a significant (or even extreme) difference between the size of the two classes. There are two types of imbalance depending on the severity:

- slight imbalance, when the distribution of observations in the training set is uneven by a small amount (such as a 4 to 6 ratio);

- severe imbalance, when the distribution of observations in the training set is uneven by a large amount (such as a 1 to 100 ratio). This is the case for bankruptcy prediction.

It must be noted that a slight imbalance is usually not concerning and the problem can be treated like a normal classification problem. However, a severe imbalance can be problematic and must be treated. In general, standard classifier learning algorithms are biased towards the majority class, as the rules that predict those observations are positively weighted in favour of the accuracy metric or the corresponding cost function. At the same time, the specific rules that predict the minority observations can be ignored (treating them as noise) in exchange for more general rules. As a result, minority instances are misclassified more often than those of the majority class. In order to correctly distinguish the minority class, a large number of techniques have been developed, however for the purpose of this thesis, two of them have been used:

- Data level approaches, which aim at rebalancing the class distribution by resampling the data space. This is a preprocessing step which decreases the imbalance without a modification of the learning algorithm. Two common data-level techniques are oversampling and undersampling. Their aim is to fix the class distribution of the dataset by changing the ratio between the different classes. These two are opposite but rather similar and simple techniques, however, more complex techniques like SMOTE (Synthetic Minority Oversampling Technique) exist. Both oversampling and undersampling introduce a bias to select more samples from one class than from another to compensate for the imbalance present in the dataset.

- Ensemble-based methods, which combine the use of an ensemble learning algorithm (such as the random forest classifier) with the abovementioned data level approach. This hybrid method adjusts the data first through rebalancing and then trains the classifier (see section 3.4.3).

Talking about data level approaches, the literature shows that applying a preprocessing step in order to rebalance the class distribution is often an effective way to deal with imbalanced datasets[7][8] [9]. In addition, these techniques have the advantage of being independent of the classifier. In the next subparagraphs, the main techniques used to ”adjust” imbalanced datasets are presented.

Random undersampling

Random undersampling refers to the non-heuristic technique used to balance class distribution by randomly eliminating instances of the majority class (figure 3.1 illustrates the procedure). It is one of the earliest rebalancing techniques and it is the main alternative to oversampling.

It must be noted that random undersampling is often critiqued for removing data that could potentially be important for the induction process. For these reasons, undersampling is applied less frequently and its uses are restricted to being practical and saving resource costs. In fact, as we are in the ”Big Data” era, undersampling is now applied more. Even if it is true that large sample size is needed to obtain valid statistical conclusions, the data must be prepared before using it. The preprocessing of data usually requires a significant human component and is specific to the task, which makes it time-consuming and expensive.

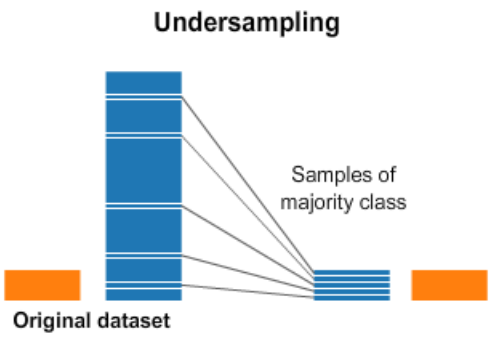

Figure 3.1 – How undersampling works. Source: Medium.com

Random oversampling

Oppositely, random oversampling is another non-heuristic approach that attempts at balancing the class distribution through the random replication of minority class instances (as demonstrated in figure 3.2).

Likewise, this is one of the earliest resampling approaches and proved to obtain good results in a low response time[5]. This method is usually preferred to undersampling, especially when the ”detailed” data has yet to be collected. One problem of oversampling is that it could increase the likelihood of overfitting, as it increases the minority class by making exact copies of the minority instances. This would cause the classifier to use apparently accurate rules when in reality these cover replicated examples. Basically, the model learns the noise in the training data to an extent that it negatively impacts the performance with the new ”rebalanced” dataset. This is problematic essentially for two reasons:

- noise and random fluctuations are captured by the training data and are learned as concepts by the model; and
- these concepts do not apply to newly recorded data, which results in a deteriorated ability of the model to generalize and decreases its accuracy.

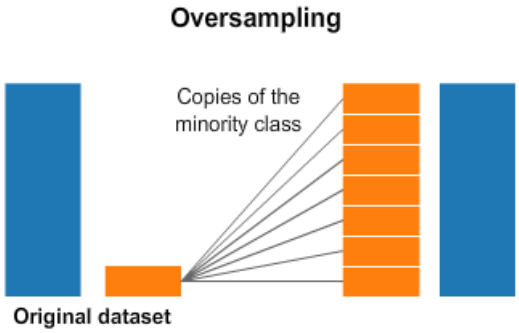

Figure 3.2 – How oversampling works. Source: Medium.com

SMOTE

Another common oversampling technique is the SMOTE approach (Synthetic Minority Oversampling Technique)[6]. The main difference with random oversampling is that SMOTE introduces synthetic examples instead of simply replicating the existing minority class examples. The SMOTE procedure is focused on the ”feature space” (rather than

the ”data space”) because the new synthetic examples are created through interpolation of many positive instances that lie together. Figure 3.3 shows a graphical example of the creation of new synthetic data points using SMOTE. First of all, the total oversampling amount N is set, in order to equalize the classes. Then, a positive example x_i is selected randomly from the training set as a basis to create the new synthetic data points. After that, its k-nearest neighbours (4 in this example) are obtained and N of these K nearest neighbours are used to compute the new synthetic examples via interpolation (data points r_1 to r_4). In order to do so, the difference between the specific feature vector and each neighbour is taken and then multiplied by a random number in the range $[0, 1]$ and ultimately added to the previous feature vector. This creates the random point along the line segments among the features.

One of the benefits of the SMOTE is that the synthetic data points created are not exact replicas of existing data points but are similar enough to known observations in the minority class. This is a more complex oversampling technique (with respect to random oversampling), however, it is a superior option.

In any case, the approach is not perfect and presents some weaknesses that have motivated academics to develop many extensions to the SMOTE algorithm[15]. SMOTE has been critiqued mainly because it generates noisy points and many of the new synthetic instances are created in the same direction, producing many borderline examples which are easily misclassified.

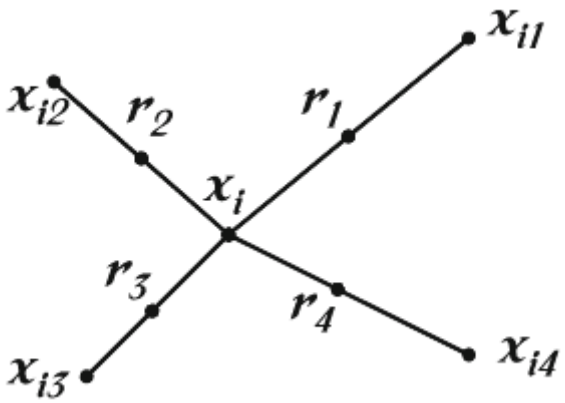

Figure 3.3 – Graphical example of the creation of new synthetic data points via SMOTE. Source: Fernández, Alberto, et al. Learning from imbalanced data sets. Vol. 10. Berlin: Springer, 2018.

3.3.3 Splitting the data into training set and test set

In order to perform the analysis on the dataset, machine learning algorithms are applied using the train-test split procedure. The train-test split procedure is used to estimate the performance of machine learning models to make predictions on the test data. It is widely used for any supervised learning algorithm, especially for classification and regression problems. Essentially, the method splits the dataset into two subsets:

- the training set, which is used to fit the model and learn the patterns of the algorithm; and
- the test set, which is used to evaluate the model that was previously fit in the training data.

The aim of the train-test split procedure is to fit the machine learning model on available data with known inputs and outputs to then make predictions on a new set (which is not used to train the model) where we do not have the expected output or target values. Samples from the original training dataset are split into two subsets using random selection to ensure that they are representative of the original dataset.

This procedure is optimal when there is a sufficiently large set of data to work with. The size of the dataset is specific to each predictive modelling problem, however, there needs to be enough to make the two subsets suitable representations of the problem domain. This means that there must be enough observations to cover all common cases and most uncommon cases. On the other hand, a small dataset could be very problematic since the training set will not be able to learn an effective mapping of inputs to outputs and the test set will not have enough observations to precisely evaluate the model performance. As a result, the evaluation could be overly optimistic or overly pessimistic. Another problem could be an imbalanced dataset (see the previous subsection), which needs to be rebalanced before applying the split to avoid poor modelling performance. In this case, the modified dataset (for example after applying SMOTE) is used only to train the model (hence the training set), while the test set performs prediction on the original imbalanced dataset. The test set remains untouched because it preserves the dataset's original distribution and in this way, it can correctly offer a valid approximation of the model with the imbalanced data. On the other hand, if the whole dataset is first resampled and then split, this will lead to an overestimation of the model since the

distribution of the classes has been altered and it does not represent the initial problem anymore. Anyway, the train-test split approach is easy to use and computationally efficient. The size of the train and test sets is adjusted to account for computational costs and representativeness in both training and testing of the model, however, the majority of the dataset is usually dedicated to the train set in order to train the model effectively, while the remaining part is dedicated to the test set (80:20 and 67:33 are usual train-test splits).

3.4 The statistical models used

The analysis is conducted using three models: the logistic regression, the LASSO regression, and the random forest. They are all statistical models and powerful classifiers operating in the supervised learning field that have proved to be well suited for the classification of imbalanced datasets. In the next paragraphs, the models in question will be presented from a theoretical viewpoint.

3.4.1 Logistic regression

The logistic regression is a statistical model used for classification and predictive analyses. Essentially, the model links the probability π of the event to a set $x = (x_1, x_2, ..., x_{p-1})$ of explanatory variables. The response variable Y takes the form of a Bernoulli random variable, whose probability of success $\pi(x)$ depends on the covariates X. If we define as $\eta(x)$ a linear combination of covariates like:

$\eta(x) = \beta_0 + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1}$

and define the logistic function as:

$\ell(\eta) = \frac{e^\eta}{1+e^\eta}$

the logistic regression model has the form:

$\pi(x) = \ell(\eta(x)) = \frac{e^{\eta(x)}}{1+e^{\eta(x)}}$

In the context of binary classification, the logistic function transforms the covariates and assigns a score in the range $[0, 1]$ to each observation. The score determines whether the event happens or not according to a threshold. By default, the threshold is set at 0.5, so if the function assigns a score above the threshold value to the observation, the instance is classified as an occurrence of the event (it must be noted that even values equal

to the threshold are categorized as occurrences). Vice versa, if the function assigns a score below 0.5, the observation is classified as "non-occurrence". Clearly, the threshold can be adjusted and this can be particularly helpful when the model has trouble recognizing one or more classes.

The standard logistic function is a sigmoid curve (S-shaped) that takes any real number and transforms it into a number in the range $[0; 1]$, which can be interpreted as a probability (figure 3.4 shows the logistic function).

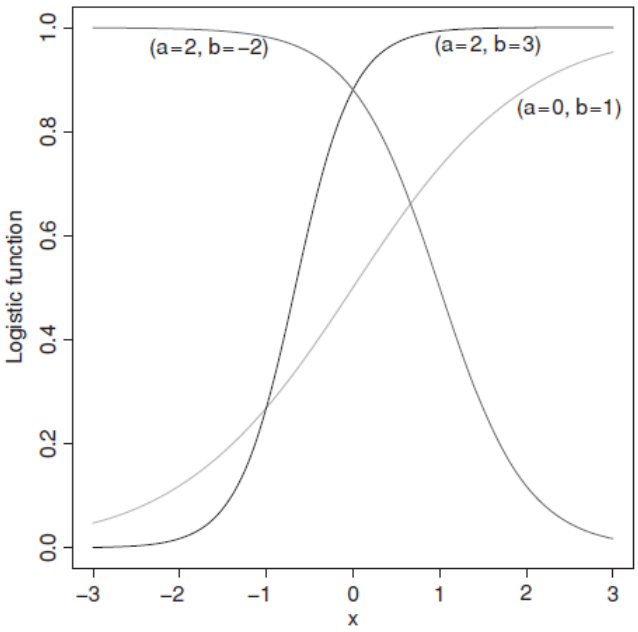

Figure 3.4 – Logistic functions with different combinations of (β_0, β_1) when $\eta(x) = \beta_0 + \beta_1 x$. Source: Azzalini, Adelchi, and Bruno Scarpa. Data analysis and data mining: An introduction. OUP USA, 2012.

There are three types of logistic models, according to the categorical response:

- binary logistic regression: in this model, the dependent variable is a dummy and can take only values 0 and 1. This is the most common form of the model and is one of the most famous binary classifiers;
- multinomial logistic regression: here, the dependent variable takes three or more variables, with no specified order (such as film genres);
- ordinal logistic regression: similar to the multinomial model, the dependent variable can take 3 or more outcomes, however, they have a defined order (such as a grading scale from 0 to 5).

The logistic regression differs in a number of ways from the linear regression. First of all, when using linear regression for classification, the model treats the categories as numbers and fits the line minimizing the distance between the data points and the line, assigning scores along the best-fitted line. However, these scores cannot be treated as probabilities and a threshold cannot be used to differentiate the classes. Additionally, the linear regression fits a straight line that takes values in the range $[-\infty; \infty]$, which cannot be interpreted as probabilities since they exceed the $[0, 1]$ range. Instead, the logistic regression predicts probabilities and assigns values in the range $[0; 1]$.

3.4.2 LASSO regression and regularization

LASSO (Least Absolute Shrinkage and Selection Operator) is one of the most famous shrinkage techniques. The aim of shrinkage is to constrain or regularize or shrink the coefficient estimates of the p predictors of the model towards zero, in order to reduce significantly their variance. LASSO is the main alternative to ridge regression (the other main shrinkage method), which instead shrinks the coefficients towards zero (both penalization methods shrink coefficients according to a parameter λ, which increases the shrinkage as the parameter increases), however, it will not set any of them exactly to zero (unless $\lambda = \infty$). The ridge penalization may not affect prediction accuracy, however, it negatively affects model interpretation when the number of predictors p is large. Instead, LASSO coefficients $\hat{\beta}_\lambda^L$ minimize the quantity:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| = RSS + \lambda \sum_{j=1}^{p}|\beta_j|.$$

The formulation is similar to that of the ridge regression, and the only difference is that the LASSO $|\beta_j|$ term is substituted by β_j^2 in the ridge penalty. While ridge is said to use an ℓ_2 penalty, LASSO uses an ℓ_1 penalty. The ℓ_1 rule of a coefficient vector β is given by $||\beta||_1 = \sum |\beta_j|$.

As mentioned above, when the tuning parameter λ is sufficiently large, the ℓ_1 penalty shrinks the coefficient estimates exactly to zero, performing feature selection on the model. As a consequence, LASSO aims at obtaining sparse models (only using a subset of variables), which are more easily interpretable than those created with the ℓ_2 penalty.

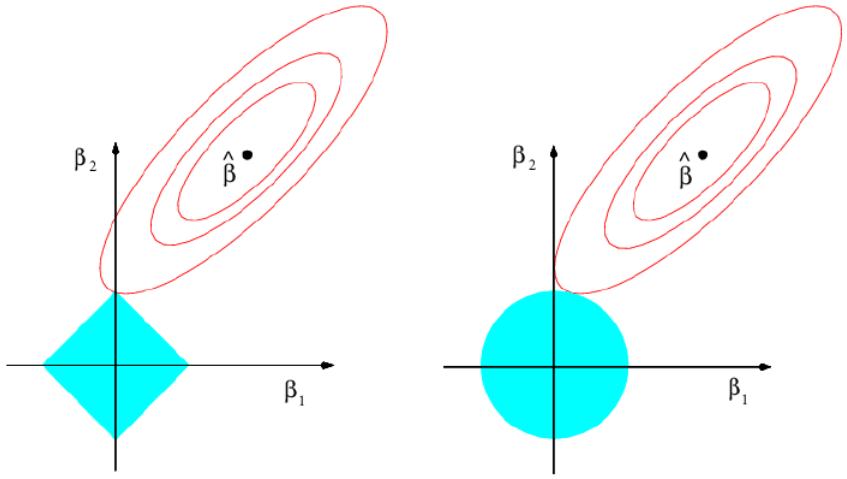

Figure 3.5 – Constraint functions for LASSO (left) and ridge (right). Source: James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.

Figure 3.5 illustrates graphically the difference between ridge and LASSO. $\hat{\beta}$ represents the least squares solution, the blue diamond and circle are respectively the LASSO and ridge constraint regions, while the red ellipses are the contours of the RSS. In the figure, the least squares estimates are outside the blue regions, meaning that they are different from the LASSO and ridge estimates. In general, decreasing the tuning parameter λ increases the constrained region, and when $\lambda = 0$, the constrained region will contain $\hat{\beta}$ and the least squares estimates will equate to those of LASSO and ridge. The red ellipses centred around the least squares are contours, and all the points in each ellipse have the same RSS. As the ellipses expand away from $\hat{\beta}$, the RSS increases. The LASSO and ridge coefficients are given by the first point of contact between the ellipse and the constraint region. However, the two types of penalties have different regions: ridge has a circle with no sharp points, meaning that the intersection will generally not occur on an axis and this causes the coefficient to be different from zero. On the other hand, the blue diamond of LASSO can often intersect with the ellipses at an axis, and every time this happens one of the coefficients will shrink to zero. This is particularly true for higher dimensions, where the intersection may happen more than once and many coefficients may be zero. However, it must be said that the choice of the penalization parameter λ is crucial and depends on the problem itself. In general, this creates a trade-off between model interpretability and information loss, with higher values of λ which create more sparse models but cause a degree of information loss.

3.4.3 Classification trees and random forests

Classification trees are supervised learning algorithms that use decision trees as predictive models to draw conclusions about a set of observations. Differently from regression trees, they are used to predict a qualitative response. In classification trees, we predict that each observation belongs to the most frequently occurring class of training instances in the region in which it belongs. When interpreting the results, we are interested both in the class prediction and the class proportion (among the training observations) of a particular terminal node region. In order to grow a tree, two criteria are often used to make binary splits: the Gini index and entropy. First of all, we define $\hat{p}_{mk}$ as the proportion of training observations in the *m*th region from the *k*th class.

The Gini index is computed as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

and can be considered as a measure of the total variance in the K classes. The Gini index depends on the values of the $\hat{p}_{mk}$'s, and if these are close to zero or one, the index is low. This metric assesses the node purity and a low G indicates that the vast majority of observations in a node belong to one class only.

Alternatively, entropy is defined as:

$$D = -\sum_{k=1}^{K} \hat{p}_{mk}\log\hat{p}_{mk}.$$

This criterion is similar numerically to the Gini index, in fact, a low value of entropy corresponds to a pure node. As $0 \leq \hat{p}_{mk} \leq 1$, then $0 \leq -\hat{p}_{mk}\log\hat{p}_{mk}$, and entropy will be low if the $\hat{p}_{mk}$ values are close to zero or one. Both of these methods are preferred to others (like the classification error rate) when evaluating the quality of a binary split since they are more sensitive to node purity.

The random forest is a tree-based statistical model that can be used both for classification and regression. Practically, the algorithm uses decision trees that represent the decision-making of the model with a branching structure (similar to a tree). In a tree, the endpoints (or leaves) of the branches represent the classes and that is where the branches stop splitting. In classification, the model uses trees as base classifiers, choosing randomly the features to be included in each model and then combining the output of the decision trees to reach a single outcome. At every node, the random forest selects randomly a small subset q of explanatory variables ($q \ll p$), which are assessed to find

their best point of subdivision according to the splitting criterion chosen (Gini index or entropy). The selection is usually made with a bagging procedure, which is usually aimed at improving prediction accuracy. Bagging is also useful to obtain the "importance" of the explanatory variables. In order to do so, first, we compute the misclassification error on the out-of-bag portion of data (all data not used for sampling). Then, we compute it again after permuting randomly the values of each covariate. Lastly, we take the average of the difference between the two misclassification errors and we divide it by the standard deviation of the difference. In this way, we obtain a measure of how an explanatory variable influences the predictions. When the set of covariates is chosen, the tree grows to maximum size but is not pruned [1]. As a result, the combination of various uncorrelated trees that are part of the random forest avoids overfitting.

There are two important tuning parameters that must be set to construct the random forest. First, the amount of q covariates selected in each node must be set and it is usually equal for all nodes. Generally, the value is chosen considering more values of q and taking the one which minimizes the error on a test set.

Second, the number of trees B composing the random forest must be set. As mentioned above, the combination of trees does not cause overfitting, so even a large value of B does not deviate much the prediction error from its minimum.

The random forest has proved to be a robust model and has some advantages over other methods of model combination. First of all, it provides a reduced risk of overfitting since the averaging of uncorrelated trees lowers the overall variance and prediction error. Second, it is a flexible method, since it can be used effectively for classification and regression. Third, prediction accuracy is similar (if not better) compared to that of boosting algorithms, however random forests are faster and require less computational power since the trees composing the forest are based on a smaller number of variables. The model is also quite simple to build since it is based on parallel computing. However, it must be noted that random forests are best suited for the analysis of large datasets and can offer interesting results only with a high number of variables.

[1] Tree-based algorithms often undergo pruning, which is a data compression technique that reduces the size of trees by removing sections that are irrelevant or redundant for classifying observations.

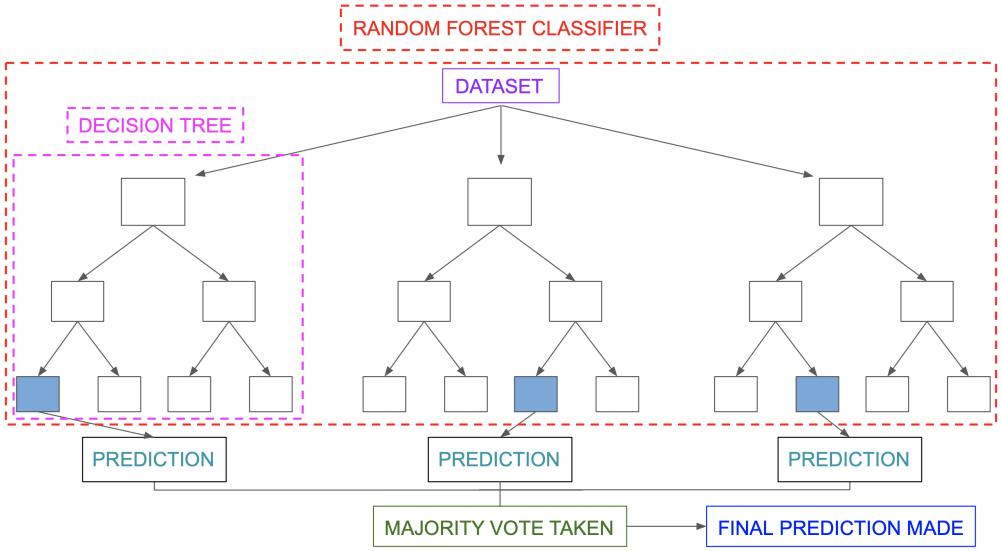

Figure 3.6 – Functioning of the random forest classifier. Source: Section.io

3.5 Performance metrics

Classification metrics are tools used to evaluate the performance of a classifier. In order to evaluate the models, some metrics have been chosen and these are the confusion matrix, the ROC curve, the ROC-AUC score, the F1 score, and other ”secondary” measures (like accuracy, precision, or recall).

3.5.1 Confusion matrix

One of the most famous tools to measure the performance of a classification algorithm is the confusion matrix. It is a particular type of contingency table with two dimensions (”actual values” and ”predicted values”) and identical sets of ”classes” in both dimensions (in the confusion matrix each combination of dimension and class forms a variable). This tool can be used for every type of classification problem and is very important to measure Recall, Precision, Specificity, Accuracy, and the ROC-AUC curve. In a binary classification example, the table has the following layout:

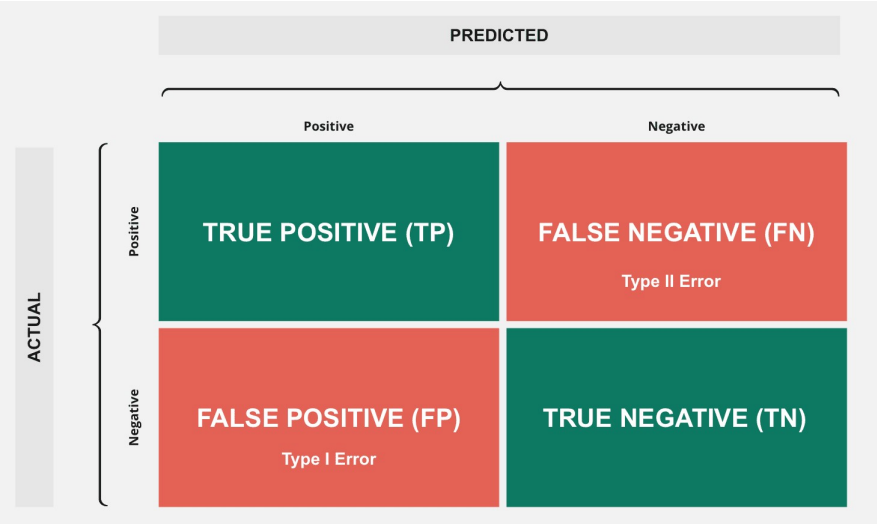

Figure 3.7 – The confusion matrix. Source: Dataaspirant

Here four categorizations of values can be identified and these are:

- True positive: the algorithm predicts positive and it is actually positive.
- True negative: the algorithm predicts negative and it is actually negative.
- False positive (Type I error): the algorithm predicts positive but in reality it is negative.
- False negative (Type II error): the algorithm predicts negative but is positive in reality.

The confusion matrix is a great indicator of how the model performs and more classification metrics can be derived from it. However, it must be noted that some of these metrics are more important than others when dealing with imbalanced data. For example, accuracy can be misleading because, in a context of high imbalance, we can still obtain a high accuracy value from the predictions of the majority class. This means that the model correctly predicts the majority class while fails to predict the minority class most of the time (which is usually the main class to look at) and this creates an illusion of the model's "efficiency" in predicting both classes. Instead, we should use the F1 score since it gives information about false positives and false negatives which are really important in imbalanced classification. Anyway, accuracy is still a good measure for balanced classification, so depending on the task and the data, some measures may be more appropriate than others.

Precision

Precision is computed as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Essentially, the metric indicates how many of our positive predictions are actually positive. The value should be as high as possible.

Recall

Recall is computed as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

It tells how many correct predictions we obtained from all the positive classes. The value should be as high as possible.

Accuracy

Accuracy is computed as:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Positives + Total\ Negatives}$$

The metric says how many correct predictions we obtained from all predictions and basically measures the "accuracy" of the main diagonal. The value should be as high as possible.

F1 Score

The F1 Score is computed as:

$$F1Score = \frac{2 * Recall * Precision}{Recall + Precision}$$

It is used to measure precision and recall at the same time and for model comparison, since it can be problematic to measure two models with high precision and low recall or vice versa. The metric uses harmonic mean instead of arithmetic mean to penalize the extreme values more.

3.5.2 ROC curve and AUC score

The ROC curve (Receiver Operating Characteristic) is a graphical plot that shows the performance of a binary classifier with changing threshold.

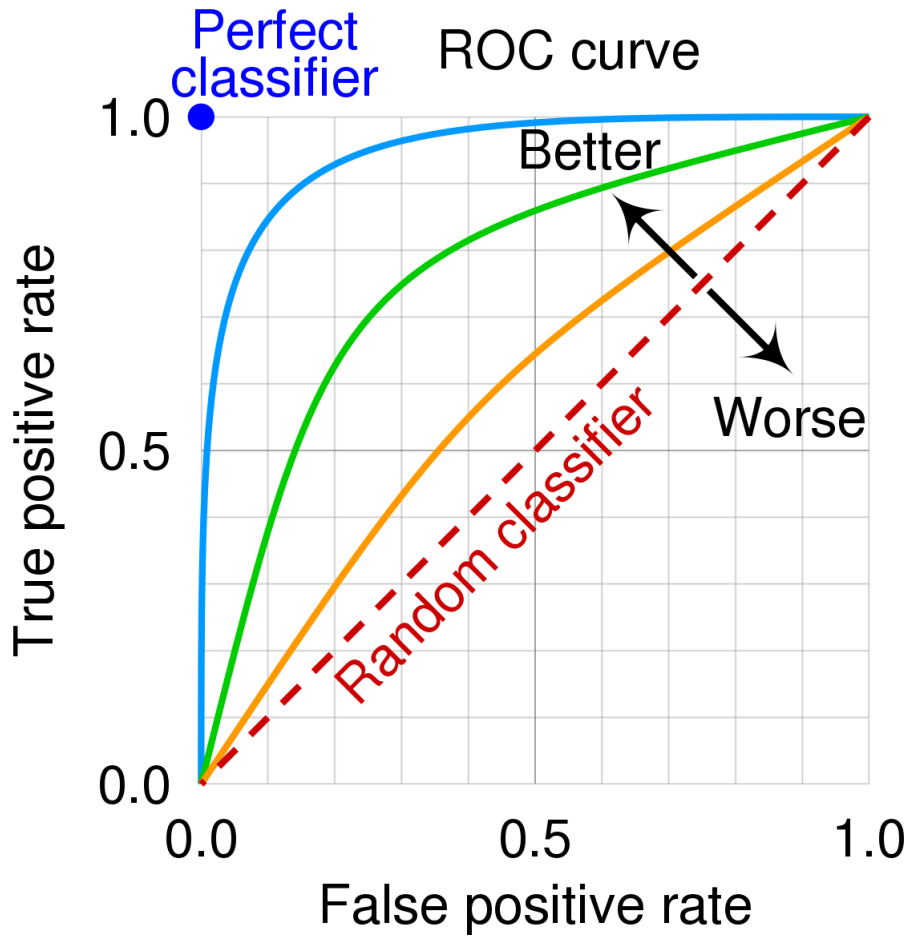

Figure 3.8 – The ROC curve. Source: Wikipedia.com

The curve is created by plotting the true positive rate (FPR, also known as sensitivity or recall) against the false positive rate (computed as $1 - FPR$) and expresses the power as a function of the Type I error of the decision rule (based on a sample of the data). As figure 3.8 shows, the slope of the curve is an indicator of the classifier's performance. If it is a straight line going from the origin to the top right corner, the classifier has a random performance level and this is the bare minimum level to consider it. While if the curve tends more to the bottom right corner, the classifier is bad and is not appropriate for the problem. As the curve bends more to the top left corner, the performance of the classifier improves, with a perfect classifier being the one forming a 90° degree angle. In real life, good classifiers are considered those in the range between the random classifier and the perfect classifier. The ROC curve is a useful indicator to compare supervised learning algorithms and select optimal ones, independently from the class distribution.

The AUC (Area Under the Curve) is the integral of the ROC curve and measures the whole two-dimensional area underneath the curve. Figure 3.9 displays it:

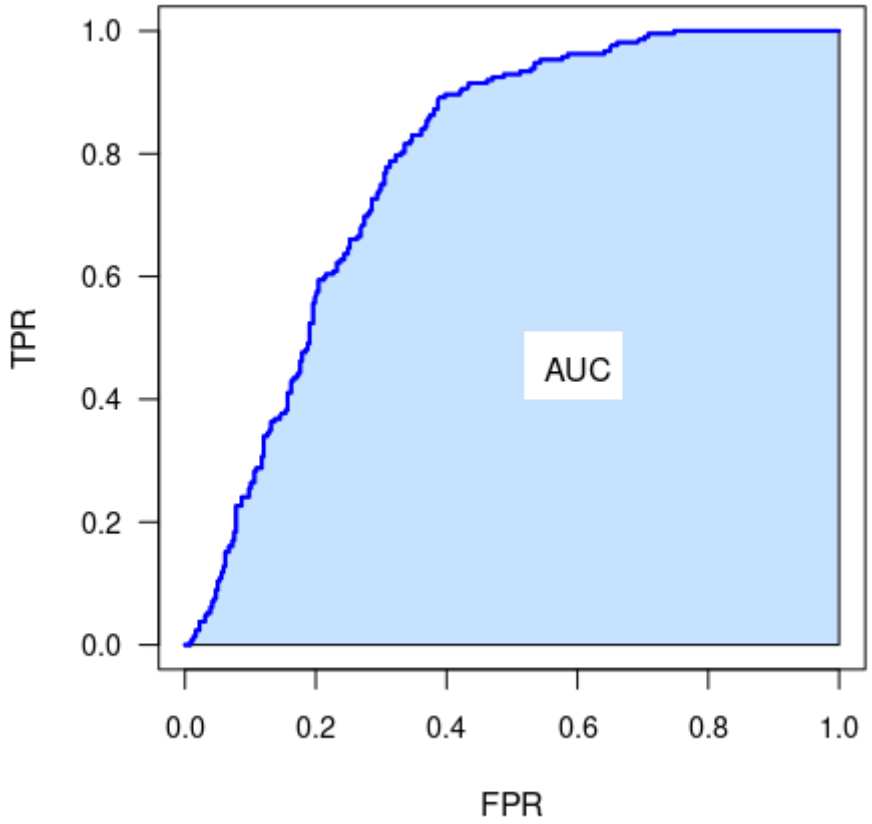

Figure 3.9 – The Area Under the Curve (AUC). Source: Wikipedia.com

AUC provides an aggregate measure of performance considering all possible classification thresholds. It can be interpreted as the probability that the model ranks a random positive observation more highly than a random negative one. It is measured through a score where 0 indicates that the model always predicts wrong while 1 means perfect prediction. Clearly, the AUC score is linked to the performance of the ROC curve and the same rules apply: good models have an AUC score higher than 0.5 and as close to 1 as possible. The two metrics (ROC and AUC) are tied together: in fact, sometimes two models may have the same AUC score but different ROC curves so it is important to look at both of them for a good evaluation.

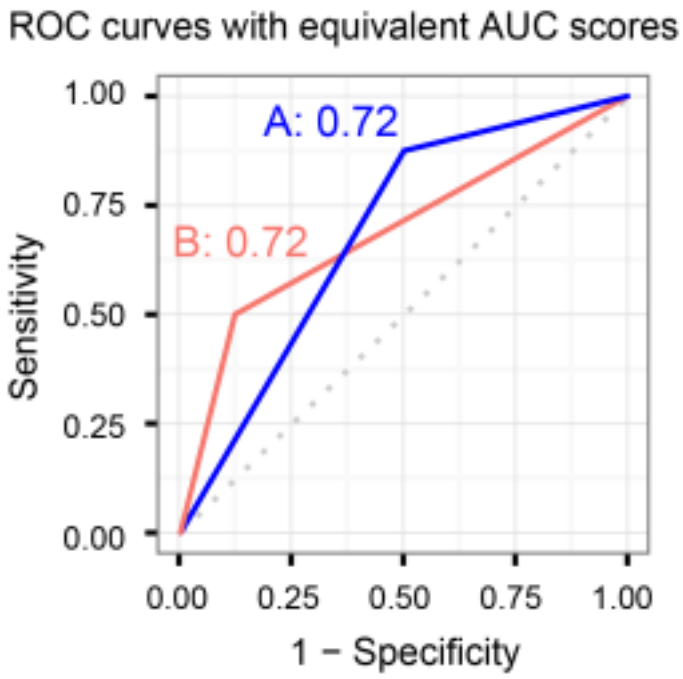

Figure 3.10 – Source: classeval.wordpress.com

ROC can be less efficient when dealing with imbalanced datasets. In order to overcome this limitation, one could look at the early retrieval area, which is a region with high specificity values in the ROC space and is useful to check the performance of the model with a small false positive rate.

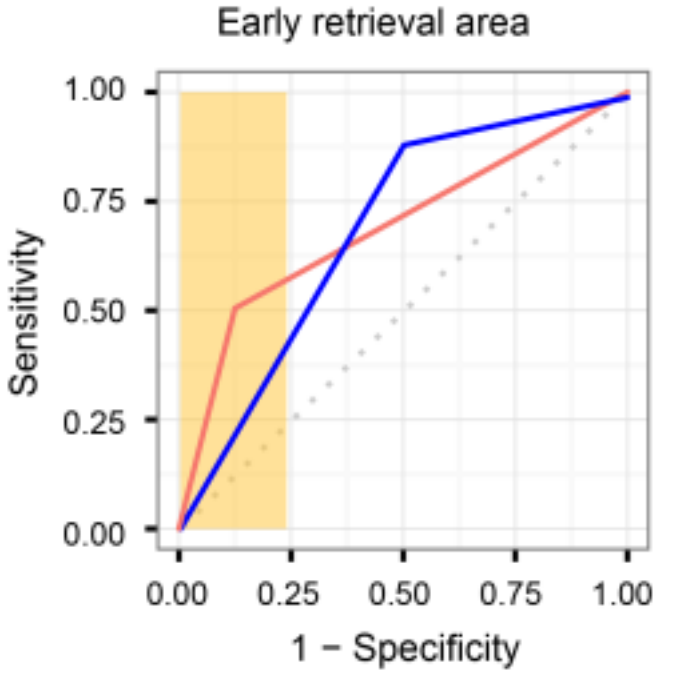

Figure 3.11 – Source: classeval.wordpress.com

4 The dataset analysis

4.1 Introduction

In this chapter, the analysis of the dataset is presented. Bankruptcy prediction is performed on a set of companies' data collected from the Taiwan Economic Journal in the period 1999-2009. The data contains 6819 firm entries and 96 variables. Here, 95 of them are features while 1 is the dependent variable.

After an initial process of data loading and cleaning, the analysis starts from the visualization of data, capturing relevant features of the dataset, which will be useful to define the strategy for data modelling. It is right in this step that the large class imbalance is noticed and in order to apply supervised learning algorithms effectively, the disproportion of the class distribution must be adjusted with a modification of the training set. After doing this, the dataset is finally ready to be modelled and the three models are fitted. Lastly, the algorithms are evaluated and compared according to popular classification metrics (confusion matrix, ROC-AUC curve, F1 score and others) and the superior model in terms of predictive power is declared.

4.1.1 Variable types

After loading the dataset and the relative libraries and functions needed, we perform a procedure of data cleaning making sure that there are no missing values and no duplicates. Right after that, the first pieces of information on the data are gathered: there are 6819 entries and 96 columns. The columns represent the variables while the entries are the different firms' information. The dataset splits the variables into two categories: 93 are numerical (*float64*), while 3 are categorical (*int64*). Due to the nature of the problem, we know that one of the categorical features is the dependent variable which categorizes each observation, however, we want first to obtain more information on the other two, namely "Liability-Assets Flag" and "Net Income Flag" (two of the variables that are also present in the Ohlson O-Score). The Liability-Assets Flag explains the proportion between assets and liabilities and takes a value of 1 if total liabilities exceed total assets, otherwise 0. Usually, assets exceed liabilities so this value is often 0 (this is also explained

by the balance sheet fundamental equation $Assets = Liabilities + Equity$, where the right-hand side constitutes the firm's financing and as a matter of fact, it is quite unusual to have no equity financing).

```
0    6811
1       8
Name:  Liability-Assets Flag, dtype: int64
 Liability-Assets Flag  Bankrupt?
0                       0            6597
                        1             214
1                       1               6
                        0               2
dtype: int64
```

Figure 4.1 – Information on the Liability-Assets Flag variable

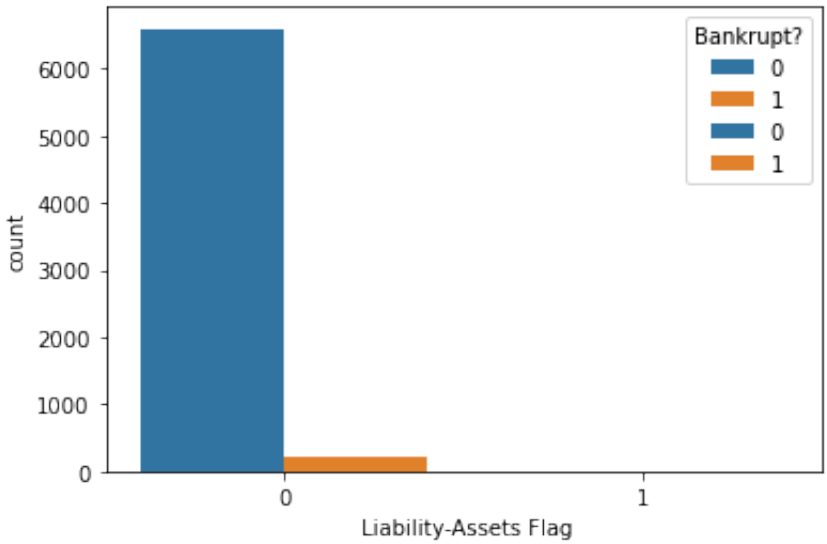

Figure 4.2 – Liability-Assets Flag distribution of values

From figures 4.1 and 4.2 we can confirm our thesis, in fact, the vast majority of firms have a value of 0 on this variable. However, it is interesting to see that the majority of bankrupt firms have liabilities exceeding assets (value 1), which means that the liability to assets ratio is high and this is a warning sign with respect to the firm's solvency.

On the other hand, the Net Income Flag denotes the firm's net income level for the last two years: if it was negative the value takes 1, otherwise, it is 0.

```
1    6819
Name:  Net Income Flag, dtype: int64
 Net Income Flag  Bankrupt?
1                 0            6599
                  1             220
dtype: int64
```

Figure 4.3 – Information on the Net Income Flag variable

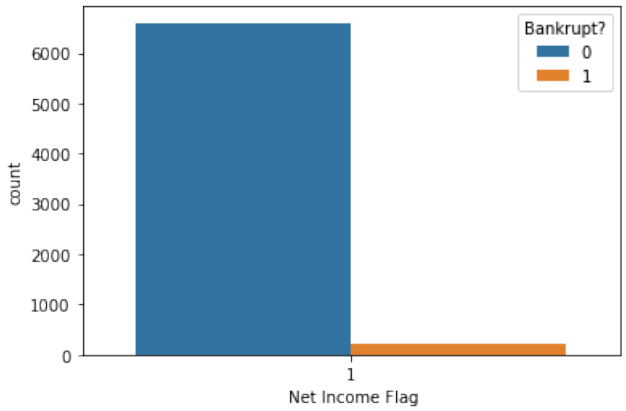

Figure 4.4 – Net Income Flag distribution of values

It is very interesting to see that the whole set of firms had negative net income in the last two years, however, 6599 of them recovered, while just 220 filed for bankruptcy.

The last categorical variable is the response, which shows that 6599 firms are financially stable, while 220 are bankrupt. This is the first sign of a high class imbalance: more than 96% of observations are categorized in one class, leaving the other with the remaining 3%. According to the reasons stated in the previous chapter, this denotes the fact that the dataset needs to be balanced before conducting model training.

```
percentage of no default is 96.77372048687491
percentage of default 3.2262795131250916

0    6599
1     220
Name: Bankrupt?, dtype: int64
```

Figure 4.5 – Information about firms' classification

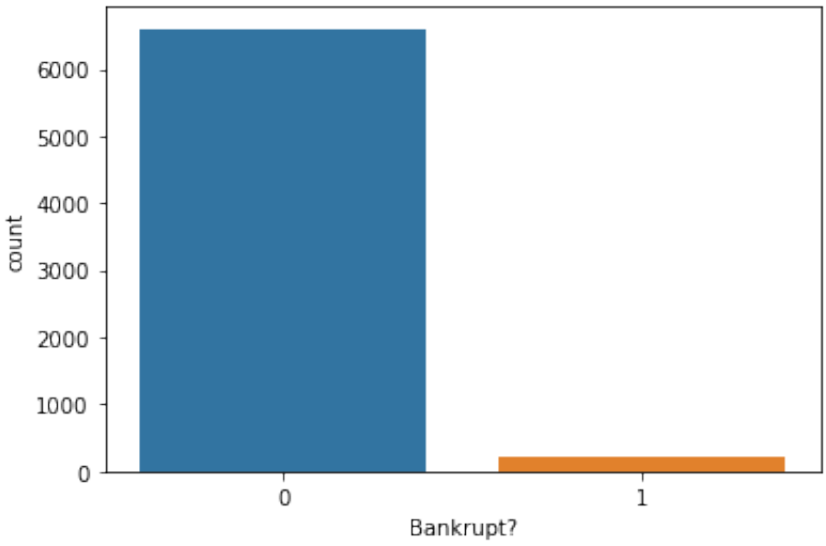

Figure 4.6 – Classification of firms

4.1.2 Exploratory Data Analysis

Now we perform the Exploratory Data Analysis in order to display as much information as possible on the variables through the use of graphical representations. First of all, let us look at the histograms of the predictors:

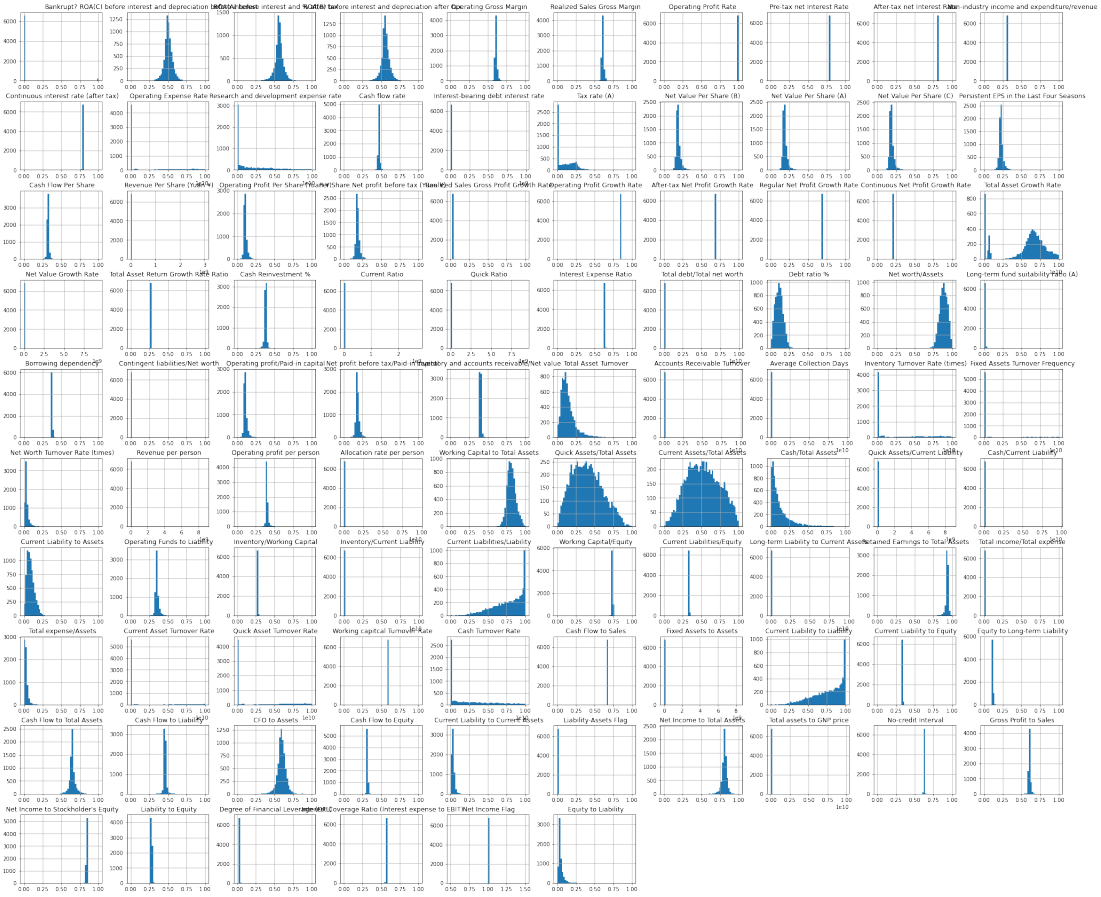

Figure 4.7 – Histograms of predictors

Looking at this figure, a good amount of variables is skewed: this means that the tail region of the variable's distribution may act as an outlier for the statistical model, adversely affecting its performance. This can limit our model selection, however, this should be verified on a case-by-case basis, depending on the degree of "anomaly" of these values and how they are related to the response. In any case, the random forest (one of our choices) is robust enough to handle outliers. All things considered, it must be noted that normalizing the data and removing outliers can be a good way to fix the problem. Subsequently, we build a Spearman correlation heatmap to check on the correlation among variables and whether they suffer from multicollinearity or not. The Spearman approach is a non-parametric measure of variables' correlation which assesses how well

the relationship between two variables can be described using a monotonic function. From the graph, we definitely see that some variables are highly correlated and we want to investigate more.

Figure 4.8 – Spearman correlation heatmap

Furthermore, we analyze the top six positively correlated features (figure 4.9). Unsurprisingly, Debt Ratio %, Current Liability To Assets and Current Liability To Current Assets are more correlated with bankrupt firms. This means that as the values of these variables rise, the number of bankrupt firms rises too. Instead, Liability to Equity and Borrowing Dependency present similar values for both classes, however, they are slightly higher for bankrupt firms. All three ratios express a high quantity of liabilities, hence a high degree of financial leverage, which is known to put firms at a higher risk of defaulting on their loans, and ultimately risking bankruptcy.

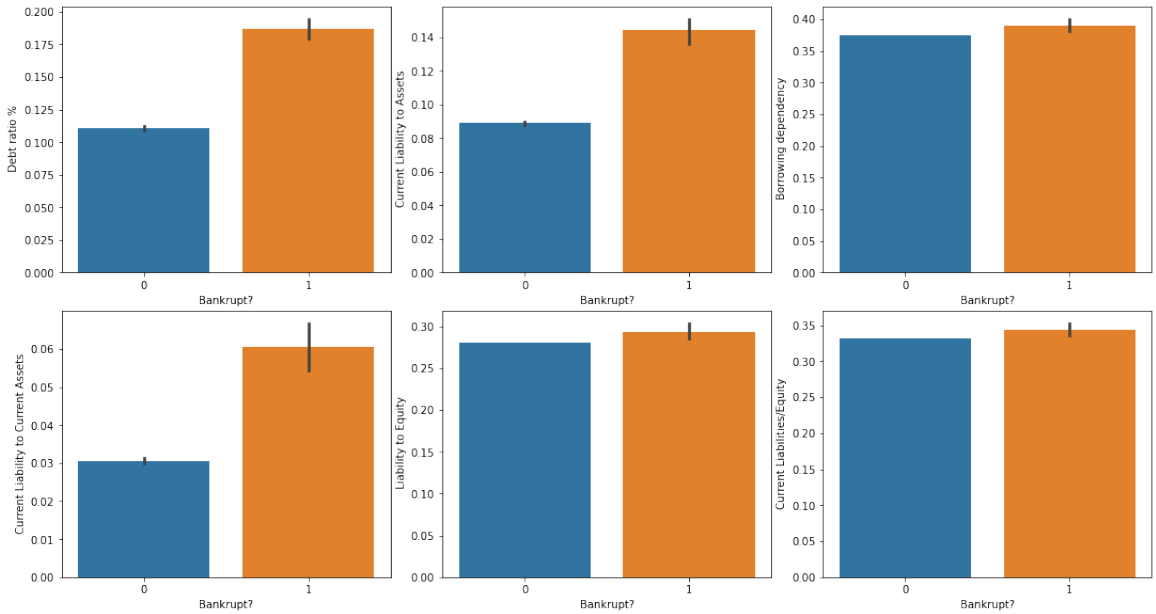

Figure 4.9 – Top six positively correlated features

To complete the Exploratory Data Analysis, we plot the distributions of the top six positively correlated variables and in each one of them, we fit the Normal Distribution to make a comparison. We see that most of them (except only for the Current Liability to Assets variable) have lighter tails than the Normal (low kurtosis) which means that there are few outliers. In addition, they also have a low standard deviation, as most of the values are centred around the mean. These distributions appear fairly symmetric.

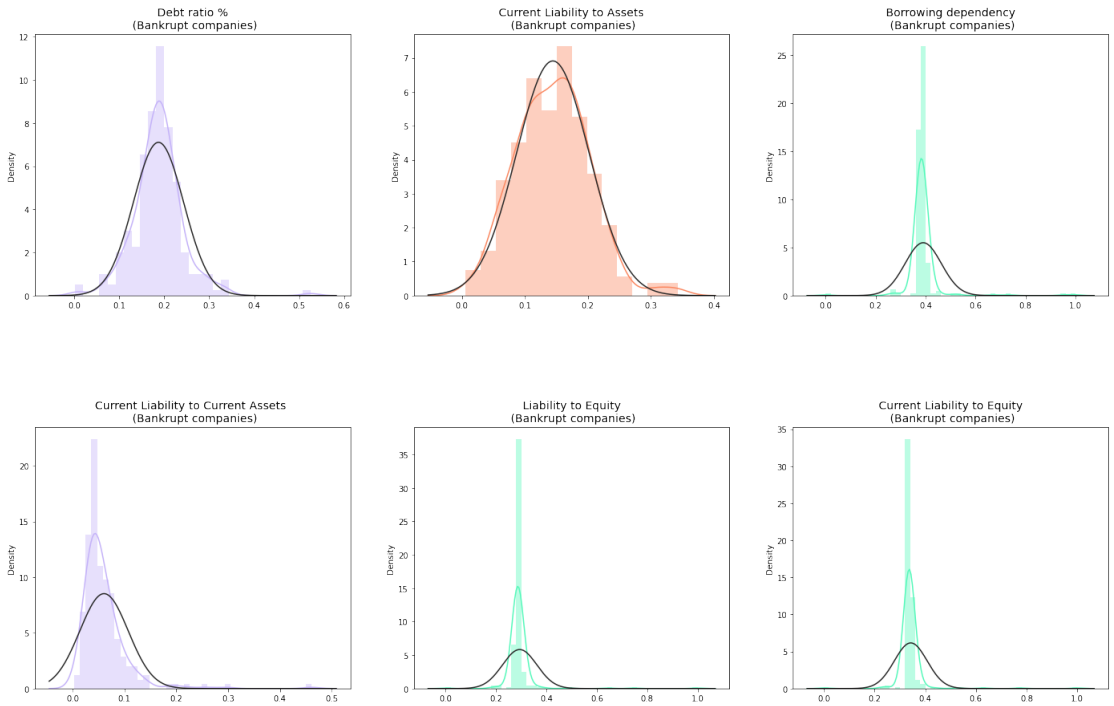

Figure 4.10 – Top six positively correlated features' distribution

4.2 Model fitting and model evaluation

In this section, the models are fitted to the dataset and evaluated to select the best classifier. From the previous analyses, a large class imbalance is noticed: 96.8% of observations fall in the ”financially stable” class and 3.2% fall in the ”bankrupt” class and, in order to obtain accurate predictions on the dataset, we first need to fix the imbalance. In order to fit a model to the data, we need to split the dataset into a training set and a test set: in the first one, the patterns of the model are learnt, while the second one is used for model testing. To avoid inaccuracy, we apply the resampling techniques only to the training set, while we leave the test set untouched. In this way, with better classes proportion, the model learns to recognize both labels in the modified training set and then is tested on a portion of the original dataset to capture the true problem. In addition, we use a technique called Stratified k-fold Cross Validation which divides randomly the set of observations into k groups (folds) of approximately the same size. The first fold is the validation set, while the method is fit on the other $k-1$ folds. Then, the mean squared error (MSE) is computed on the instances of the held-out fold. This technique is repeated k times (here $k = 5$), and each time the validation set uses a different group of observations. The stratification ensures that the original class proportion is maintained in both subsets and it is helpful to make sure that no value is either under-represented or over-represented in the two subsets, obtaining a better prediction. After these adjustments, the models can be fitted. In order to optimize the algorithms, a cross-validated search over the parameters settings is performed: the set of hyperparameters (parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning) is chosen randomly from the distribution and a score is attributed to them. The set of hyperparameters with the best score is chosen as the optimal one for the model. In the last step, the three classifiers are compared between them (using classification metrics) and the superior model is selected. The next subsections will confront each model according to the resampling technique used.

4.2.1 Random undersampling

Before starting with resampling, the dataset is split with the following proportions: 80% in the training set and 20% in the test set (equal for all resampling techniques). Then we apply random undersampling to the training set by resampling only the majority class and check if this technique offers good results.

Logistic regression

We start with the logistic regression and after fitting the model, the classification metrics are displayed. The classification report shows a good performance in terms of the majority class, however, the minority class (which is of primary importance for our analysis) has very poor results in all the parameters. An accuracy of 0.70 is decent, however, it is inflated by the good number of predictions in the majority class.

```
Logistic Regression

Evaluation Of Models

Random Model Evaluation
              precision    recall  f1-score   support

           0       0.97      0.71      0.82      1313
           1       0.06      0.51      0.11        51

    accuracy                           0.70      1364
   macro avg       0.52      0.61      0.47      1364
weighted avg       0.94      0.70      0.80      1364
```

Figure 4.11 – Classification report of the logistic regression

The confusion matrix confirms the results of the classification report. Sensitivity is 51% and specificity is 0.71%, while the type I and type II errors are respectively 378 and 25. The logistic model has a high number of false positives, and false negatives (particularly important for the analysis) are basically half of all positive predictions.

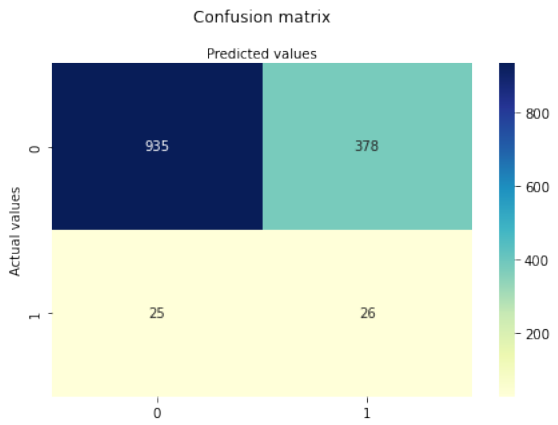

Figure 4.12 – Confusion matrix of the logistic regression

LASSO regression

Now we fit the LASSO regression and check its performance with the undersampled training data. The aim of the ℓ_1 regularization is to perform a selection of the features in order to remove those irrelevant and redundant to the model. From the classification report, we can see that the performance is equal to that of the logistic model with ℓ_2 regularization.

LASSO Regression

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	0.97	0.71	0.82	1313
1	0.06	0.51	0.11	51
accuracy			0.70	1364
macro avg	0.52	0.61	0.47	1364
weighted avg	0.94	0.70	0.79	1364

Figure 4.13 – Classification report of the LASSO regression

The only difference noticeable in the confusion matrix is that the LASSO misses three correct majority predictions with respect to the ℓ_2 regularization. Sensitivity and specificity remain unchanged, while the type I and type II errors are respectively 381 and 25.

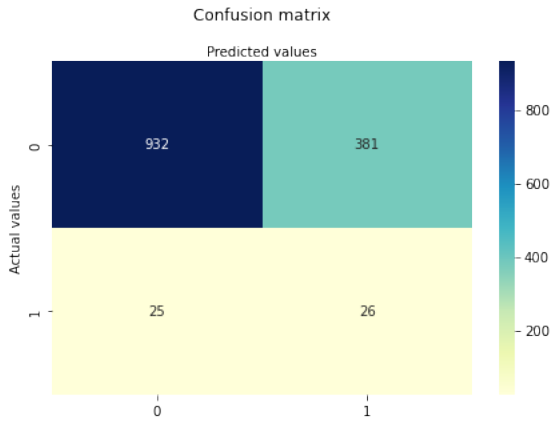

Figure 4.14 – Confusion matrix of the LASSO regression

Random forest

As the last model, we fit the random forest classifier and check its performance with undersampling. It is clear from the classification report that the performance is better with respect to the other two algorithms, however, the prediction of the minority class is still highly unreliable.

Random Forest Classifier

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	1.00	0.85	0.92	1313
1	0.21	0.98	0.34	51
accuracy			0.86	1364
macro avg	0.60	0.92	0.63	1364
weighted avg	0.97	0.86	0.90	1364

Figure 4.15 – Classification report of the random forest classifier

The confusion matrix underlines a good performance with the majority class, and the same can be said for the minority class, as the proportion of false negatives is very low with respect to true positives. The only downside of the model is a high number of false positives, which ruins in part the results. The sensitivity is 98% and the specificity is 85%, while the type I and type II errors are respectively 191 and 1.

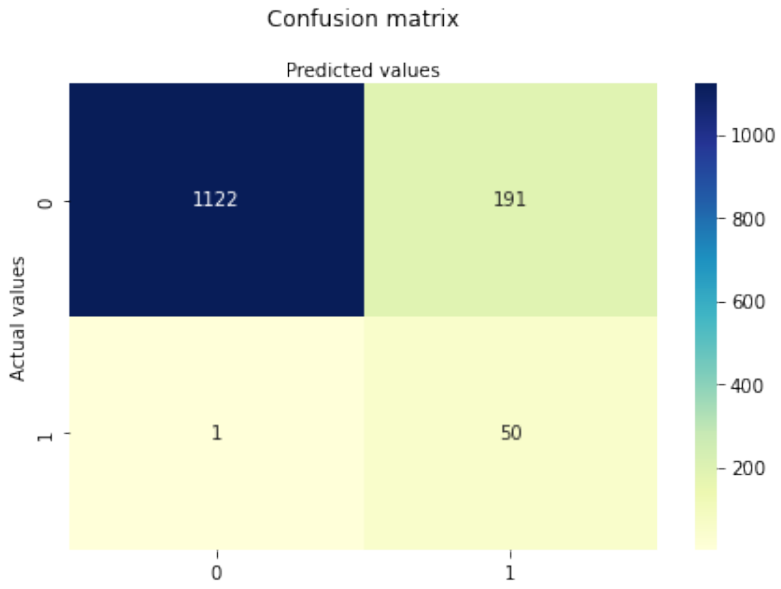

Figure 4.16 – Confusion matrix of the random forest classifier

Summary of the random undersampling application

Clearly, the random forest is the best model among the three with overall good results. However, due to a high number of false positives, precision is very low and this affects negatively the F1 score. The ROC curves confirm the results: the random forest shows overall good performance which is far superior to the other two, while the logistic regression tends to perform slightly better than LASSO. Let us see now if oversampling can improve the results, testing random oversampling first, and then SMOTE.

	Algorithm	Model Score	Precision	Recall	F1 score	ROC-AUC score
1	Random Forest Classifier	85.92%	0.21	0.98	0.34	0.92
0	Logistic Regression	70.45%	0.06	0.51	0.11	0.61
2	LASSO Regression	70.23%	0.06	0.51	0.11	0.61

Figure 4.17 – Summary of the random undersampling results

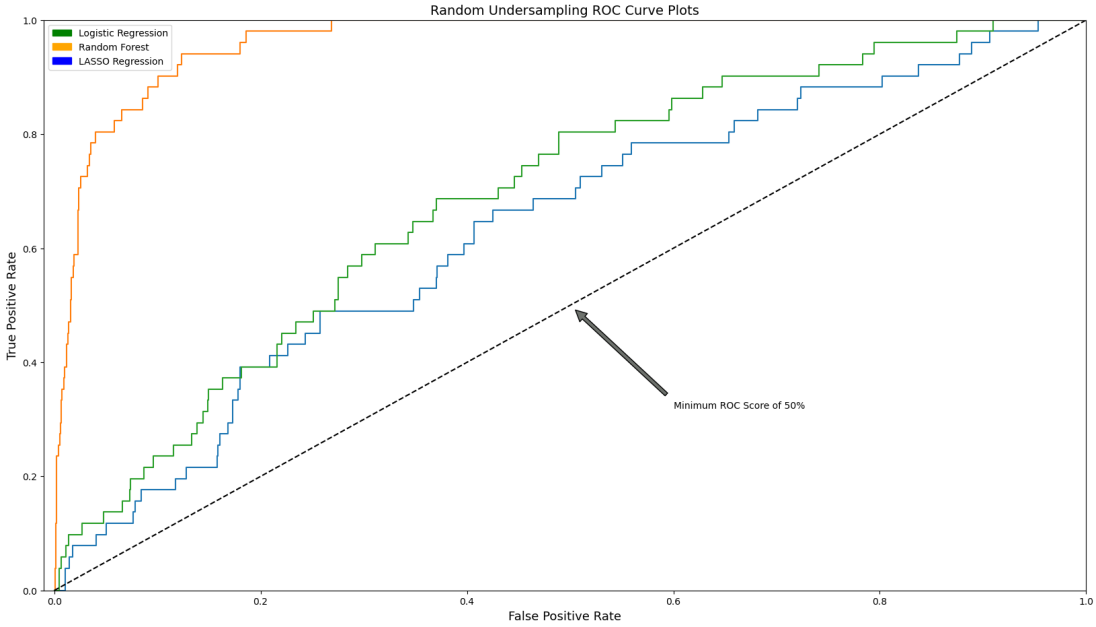

Figure 4.18 – ROC curves for random undersampling

4.2.2 Random oversampling

After evaluating the effects of random undersampling on training data, we choose to fit random oversampling. Oppositely, this technique resamples the minority class by randomly selecting examples from the minority class, with replacement, and adding them to the training dataset.

Logistic regression

Likewise, at first, we check the logistic regression. The classification report shows that there is an improvement in all parameters, however, the model is still having a hard time predicting the minority class, which is majorly important for our analysis.

Logistic Regression

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	0.97	0.76	0.85	1313
1	0.07	0.49	0.13	51
accuracy			0.75	1364
macro avg	0.52	0.62	0.49	1364
weighted avg	0.94	0.75	0.82	1364

Figure 4.19 – Classification report of the logistic regression

The confusion matrix offers a better look at these results. The improvements are attributed to a better prediction of the majority class: the algorithm obtains 60 more observations and reduces the number of false positives by the same amount. On the contrary, the minority class loses one prediction. The sensitivity is 49% and the specificity is 76%, while the type I and type II errors are respectively 320 and 26.

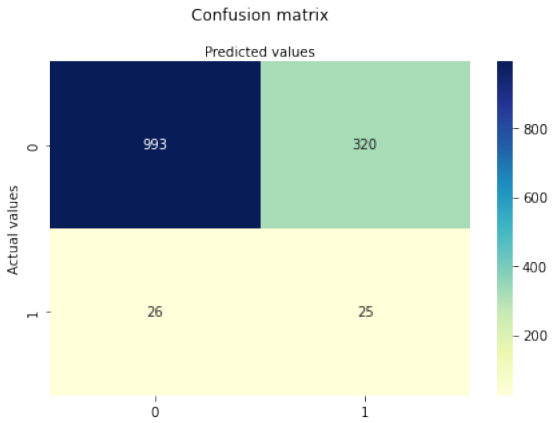

Figure 4.20 – Confusion matrix of the logistic regression

LASSO regression

For the LASSO regression, the results are similar to those of the logistic model with ℓ_2 penalty, however, the classification report highlights a marginal decrease in terms of recall, F1 score and accuracy. Anyway, the performance tends to slightly improve with respect to undersampling.

LASSO Regression

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	0.97	0.75	0.85	1313
1	0.07	0.49	0.12	51
accuracy			0.74	1364
macro avg	0.52	0.62	0.49	1364
weighted avg	0.94	0.74	0.82	1364

Figure 4.21 – Classification report of the LASSO regression

The confusion matrix looks very similar to that of logit, however here the model misclassifies a few more observations, especially in the majority class.

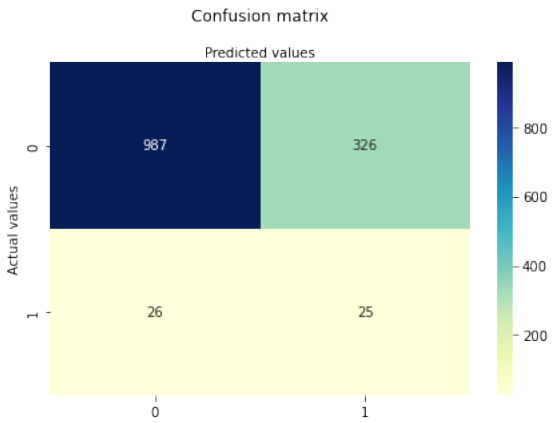

Figure 4.22 – Confusion matrix of the LASSO regression

Random forest

The true model of interest is the random forest, which proved to be the superior choice with undersampling and we now check if the minority class can be predicted better by changing the resampling technique. From the classification report, we finally obtain the results we want, with good parameters also for the minority class.

```
Random Forest Classifier

Evaluation Of Models

Random Model Evaluation
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      1313
           1       0.96      0.86      0.91        51

    accuracy                           0.99      1364
   macro avg       0.98      0.93      0.95      1364
weighted avg       0.99      0.99      0.99      1364
```

Figure 4.23 – Classification report of the random forest classifier

The confusion matrix shows that the model performs well with both classes, whereas the proportions of the two errors are quite low (2 of type I and 7 of type II. Sensitivity and specificity are 86% and 99%.

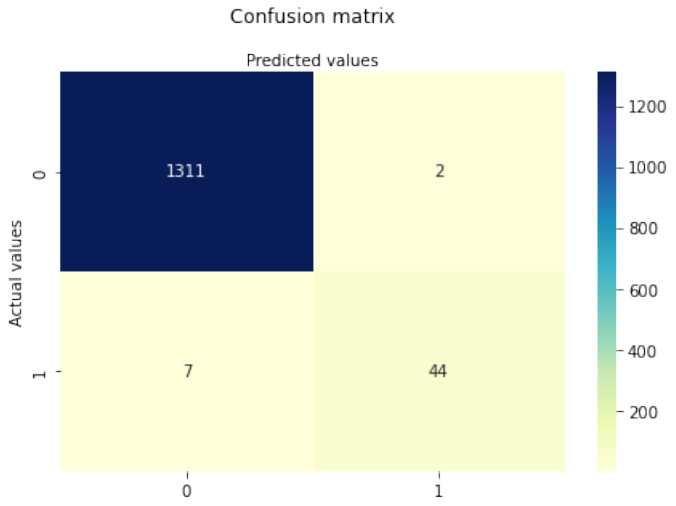

Figure 4.24 – Confusion matrix of the random forest classifier

Summary of the random oversampling application

The performance improvement that we wanted from the change of resampling technique came with random oversampling. The random forest shows very good parameters and it is clear that oversampling is better suited for this type of problem with respect to undersampling. The ROC curves show an overall better performance which is more notable for the random forest that is now close to a perfect classifier. LASSO and logistic regression have very similar curves, however, they obtained a 1-2% increase in AUC score with respect to undersampling.

Now, we try to see if with SMOTE the model can improve further, especially in terms of recall, which for this type of problem is more important than precision.

	Algorithm	Model Score	Precision	Recall	F1 score	ROC-AUC score
1	Random Forest Classifier	99.34%	0.96	0.86	0.91	0.93
0	Logistic Regression	74.63%	0.07	0.49	0.13	0.62
2	LASSO Regression	74.19%	0.07	0.49	0.12	0.62

Figure 4.25 – Summary of the random oversampling results

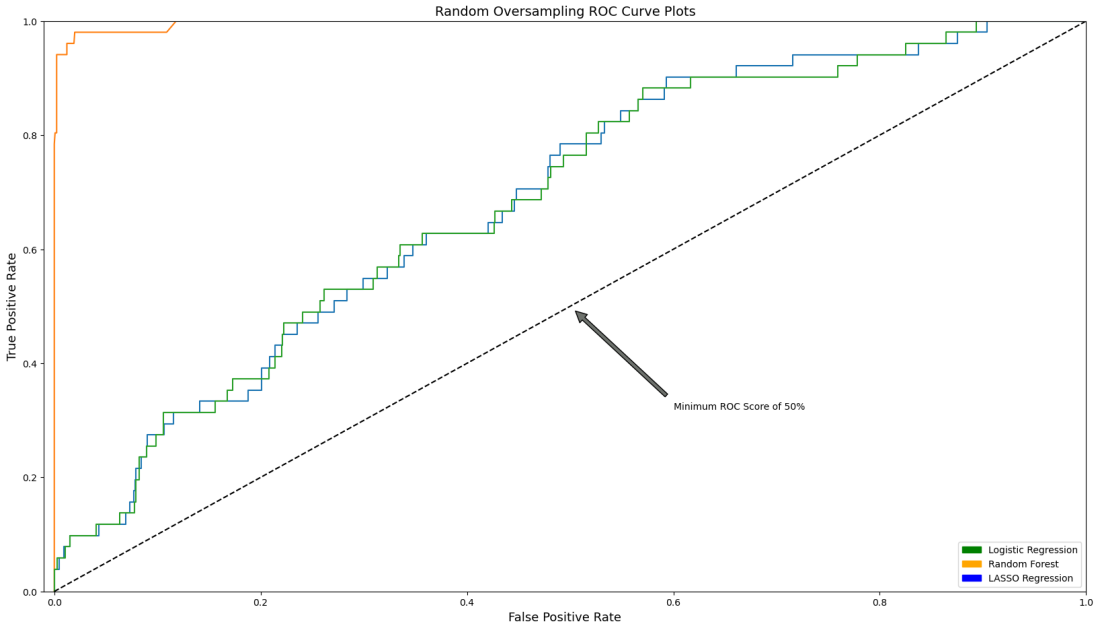

Figure 4.26 – ROC curves for random oversampling

4.2.3 SMOTE

As the last resampling technique, we apply SMOTE to the training set, which is a more complex oversampling technique than random oversampling. In the same fashion, SMOTE resamples only the minority class and we now look at further improvements to our models in order to identify the optimal resampling strategy.

Logistic regression

As usual, the logistic regression is tested first. As we can see from the classification report, the model does not perform well even with SMOTE. Precision, recall and F1 score are good for the majority class but really bad for the minority class. Accuracy is quite good, however, it is inflated by the majority class: the model can predict very well financially stable firms but hardly predicts bankrupt companies.

```
Logistic Regression

Evaluation Of Models

Random Model Evaluation
              precision    recall  f1-score   support

           0       0.98      0.81      0.88      1313
           1       0.09      0.47      0.15        51

    accuracy                           0.80      1364
   macro avg       0.53      0.64      0.52      1364
weighted avg       0.94      0.80      0.86      1364
```

Figure 4.27 – Classification report of the logistic regression

Looking at the confusion matrix, we can have a more accurate view of the predictions. The model predicts correctly 1063 stable firms and 24 bankrupt firms. The sensitivity is 47% and the specificity is 81%, while the type I and type II errors are respectively 250 and 27. Overall, the performance is better than the other two techniques, however, this only affects the majority class.

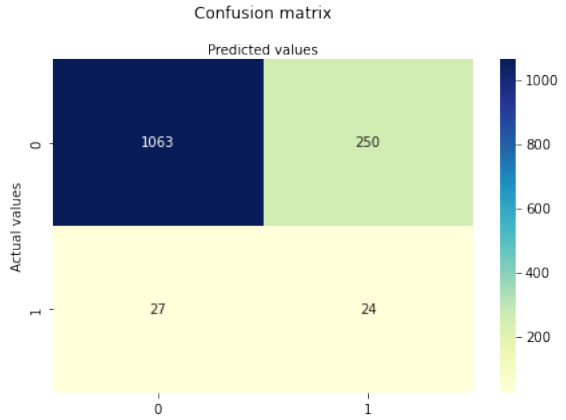

Figure 4.28 – Confusion matrix of the logistic regression

LASSO regression

As previously, the second model is the LASSO regression. The classification report is not much different from that of the logistic regression, in fact, the only change is that the model obtains 1% more in the F1 score of the majority class, however, everything else remains the same.

LASSO Regression

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	0.98	0.81	0.89	1313
1	0.09	0.47	0.15	51
accuracy			0.80	1364
macro avg	0.53	0.64	0.52	1364
weighted avg	0.94	0.80	0.86	1364

Figure 4.29 – Classification report of the LASSO regression

The same goes for the confusion matrix, the algorithm can predict only one observation in the majority class more than the logistic regression.

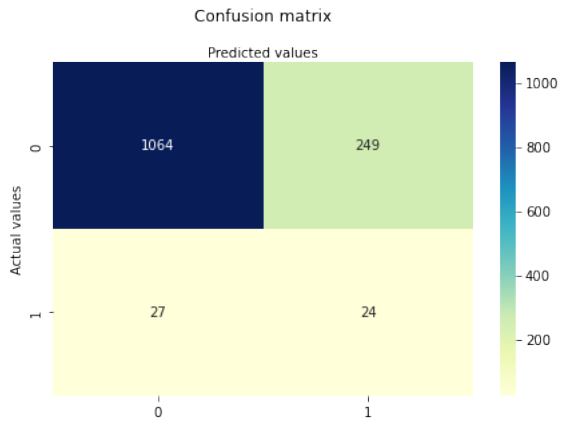

Figure 4.30 – Confusion matrix of the LASSO regression

Random forest

Finally, we fit the random forest classifier to the modified data. Here the situation is totally different, the classification report highlights that the model can predict the majority class almost perfectly, maintaining good precision, accuracy, recall and F1 score also for the minority class.

Random Forest Classifier

Evaluation Of Models

Random Model Evaluation

	precision	recall	f1-score	support
0	1.00	0.99	0.99	1313
1	0.75	0.98	0.85	51
accuracy			0.99	1364
macro avg	0.87	0.98	0.92	1364
weighted avg	0.99	0.99	0.99	1364

Figure 4.31 – Classification report of the random forest classifier

Looking at the confusion matrix, the model correctly predicts 1296 stable firms and 50 bankrupt companies. The sensitivity is 98% and the specificity is 99%, while the type I and type II errors are respectively 17 and 1. Through SMOTE, we managed to obtain a 12% increase in sensitivity, which is what we were looking for. It is true that SMOTE has more false positives than random oversampling, however, false negatives are highly reduced and the number of true positives has increased, which are good trade-offs for a problem of imbalanced classification like this.

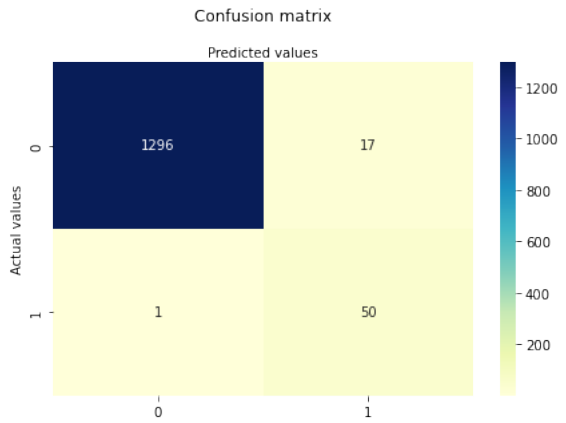

Figure 4.32 – Confusion matrix of the random forest classifier

Summary of the SMOTE application

The following table summarizes the results. The random forest is undoubtedly the best classifier among the three, considering all metrics. The ROC-AUC score increased, while the F1 score decreased slightly with respect to random oversampling, even if it reaches a good value anyways. This is due to the precision-recall trade-off and on the matter of bankruptcy prediction, we prefer higher recall than higher precision, since we give more importance to false negatives than to false positives. The ROC curves further improve and this can be proved by the AUC scores which register a 0.98 for the random forest, while a 0.64 for both the logistic and the LASSO.

	Algorithm	Model Score	Precision	Recall	F1 score	ROC-AUC score
1	Random Forest Classifier	98.68%	0.75	0.98	0.85	0.98
0	Logistic Regression	79.69%	0.09	0.47	0.15	0.64
2	LASSO Regression	79.77%	0.09	0.47	0.15	0.64

Figure 4.33 – Summary of the SMOTE oversampling results

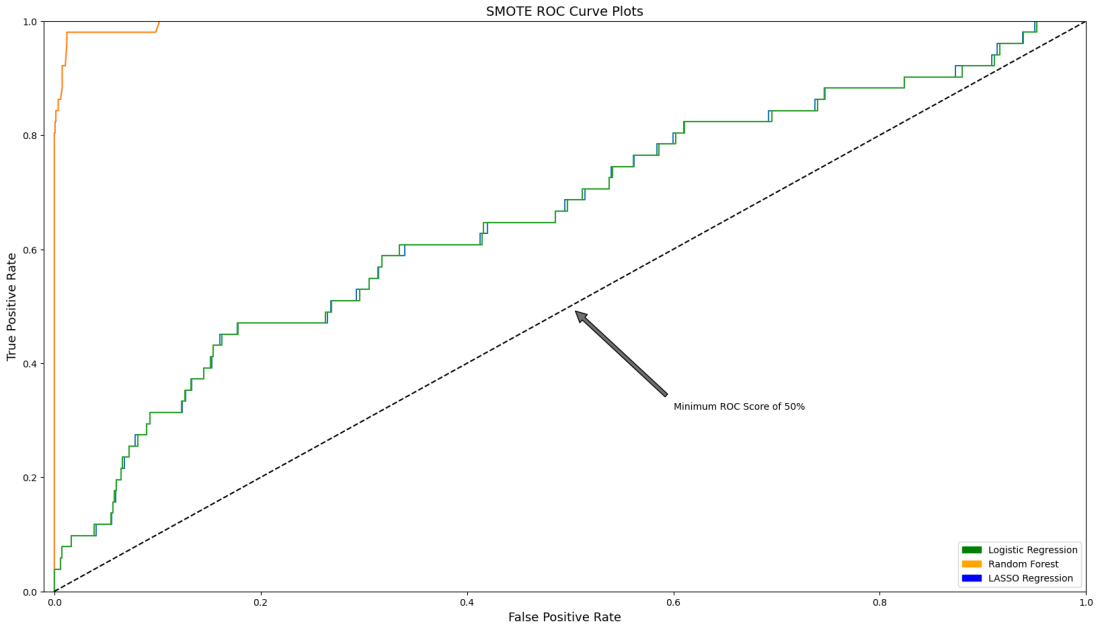

Figure 4.34 – ROC curves for SMOTE

4.2.4 Random forest feature selection

The results of the analysis show that the random forest classifier trained on resampled data using SMOTE is the superior choice for this problem of bankruptcy prediction. We now try to perform feature selection to see if the algorithm can perform even better, by eliminating redundant and irrelevant variables. We do this by selecting the $k = 81$ features that receive the highest score according to the Analysis of Variance (ANOVA) F-value. ANOVA uses F-tests to assess statistically if the means of three or more groups are different.

The classification report shows that the model performs slightly worse with less features, especially hurting the minority class prediction. This is reflected both in precision and recall for the ”bankrupt” class and ultimately on the F1 score.

```
Random Forest Classifier

Evaluation Of Models

Random Model Evaluation
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      1313
           1       0.73      0.94      0.82        51

    accuracy                           0.98      1364
   macro avg       0.86      0.96      0.91      1364
weighted avg       0.99      0.98      0.99      1364
```

Figure 4.35 – Classification report of the random forest classifier with feature selection

The confusion matrix displays these changes, which are not massive and decrease the

predictive power of the model only by a few observations, however, this underlines the fact that feature selection actually hurts the model performance.

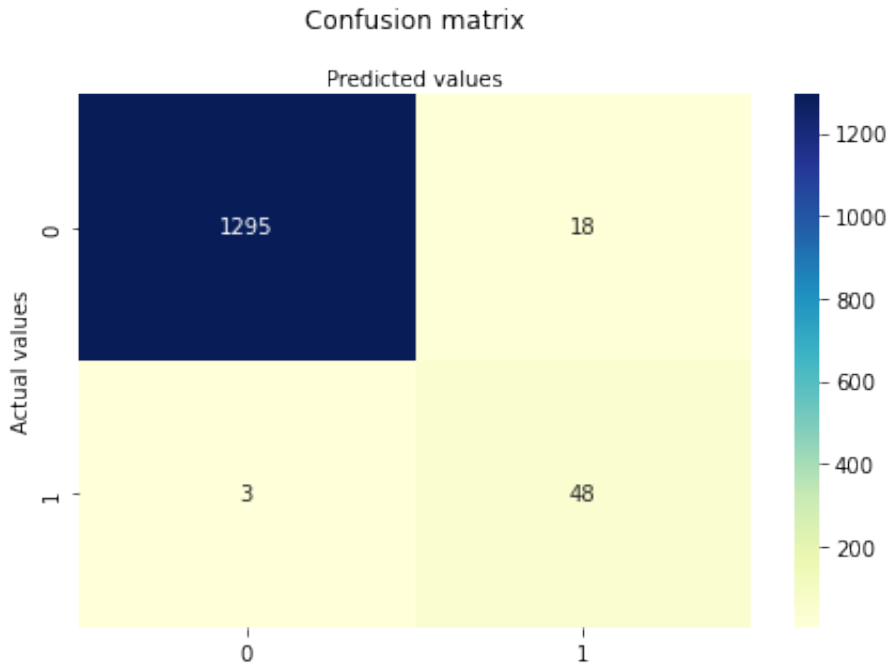

Figure 4.36 – Confusion matrix of the random forest classifier with feature selection

All things considered, feature selection actually hurts the performance of the model, especially the minority class, which is of particular interest for this analysis. On top of this, we choose to maintain all the variables as we obtain better predictive power.

4.2.5 Exploring the selected model

From the analysis, we select the random forest classifier with SMOTE oversampling on training data as the optimal model for this problem of bankruptcy prediction. As the last thing, we obtain some more information about the covariates and their relation with the response.

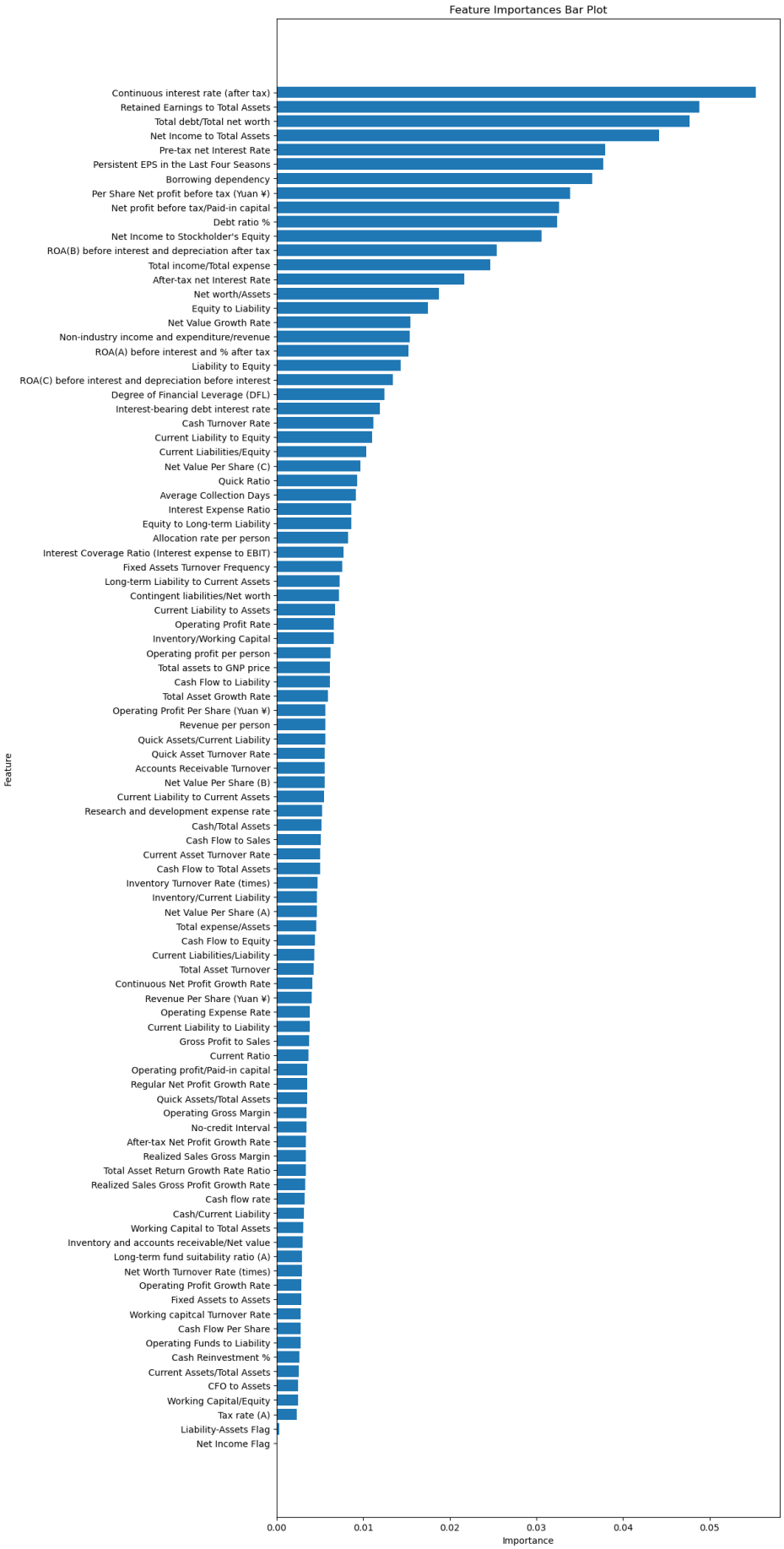

Figure 4.37 – Feature importance plot

In figure 4.37 we show the features' importance of our model. Feature importance is computed as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. Clearly, the higher the value, the more important the feature. We can see that the five most relevant features are: "Continuous interest rate (after tax)", "Retained Earnings to Total Assets", "Total Debt/Total net worth", "Net Income to Total Assets" and "Pre-Tax net Interest Rate". These five measures account for interest rates, profit for the period and accumulated earnings, and debt, which are three important areas for the firm's activity and financial health. Retained earnings to total assets shows the proportion of total assets that are covered by retained earnings and essentially measures the management's intention to use debt or new shares to invest in the company's assets. Retained earnings capture the accumulated profit or loss from the beginning of business until the reporting date and low (or even negative) values for this measure negatively impact the firm's ability to cover its assets, which is another sign of weakness. Clearly, total debt to total net worth is another important feature and high degrees of leverage increase the firm's overall risk of bankruptcy. Net income to total assets (also known as Return On Assets) shows the firm's profitability in relation to its assets and is one measure of financial strength. If a company has a low ROA, it shows that the firm is not efficient at using its assets to generate profit. This is a sign of weakness and can increase the risk of bankruptcy as well.

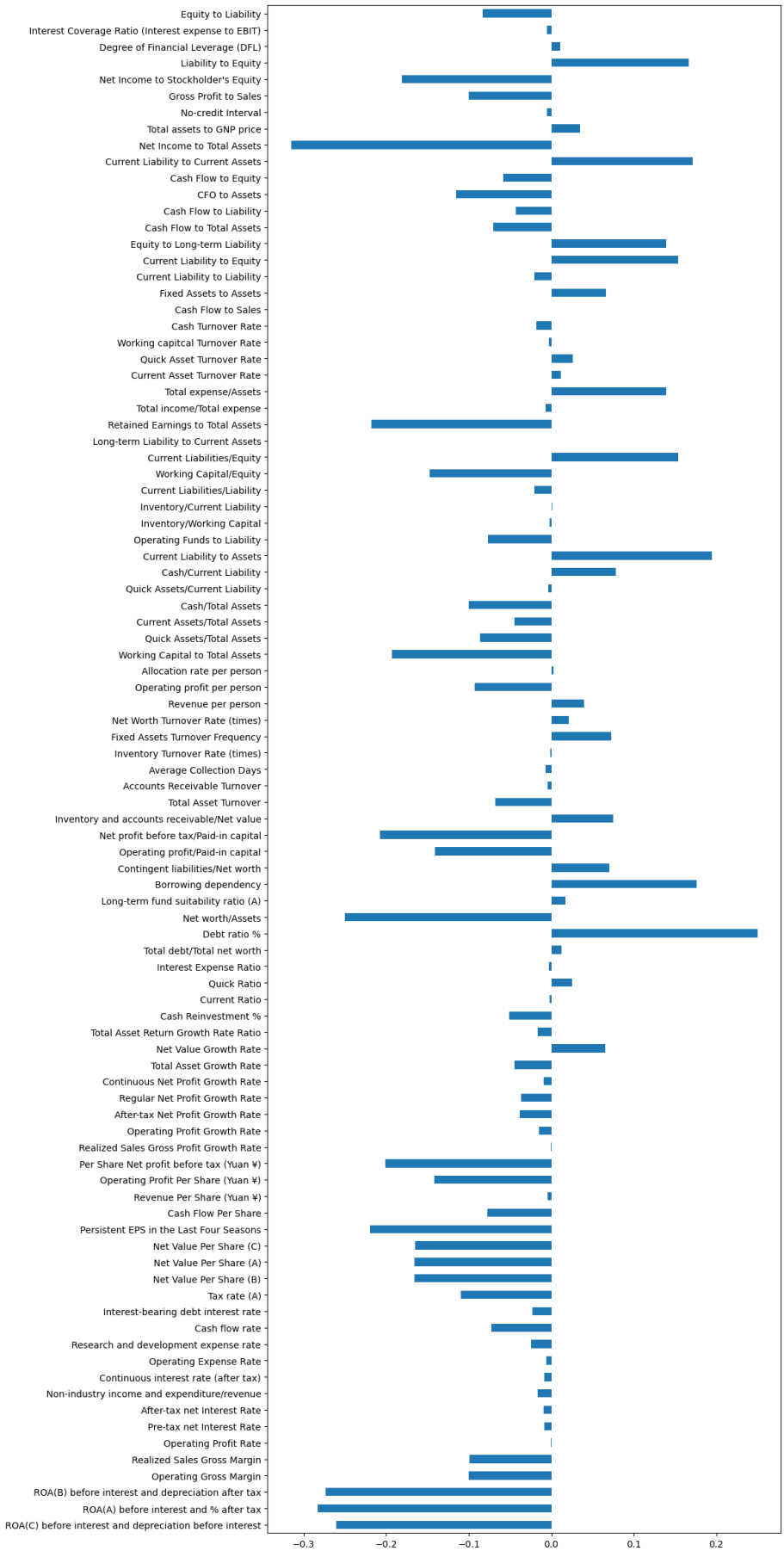

Figure 4.38 – Point biserial correlation plot

In figure 4.38 the point biserial correlation plot is displayed. We use this type of plot in order to measure the continuous covariates' correlation with the qualitative response. We can see that the "Retained Earnings to Total Assets" and "Net Income to Total Assets" features (two of the most important features of our model) are among the most negatively correlated with the response, meaning that the higher these values, the less the number of bankruptcies. In general, all ratios related to profits and the firm's efficiency are negatively correlated with the response, which highlights the importance of net income for the financial stability of a firm. On the opposite, ratios referring to liabilities are positively correlated with y, meaning that the higher the number of liabilities (and leverage), the higher the risk of the firm going bankrupt. "Debt Ratio %" is the variable with the highest positive correlation with the response, followed by "Current Liability to Assets" and "Current Liability to Current Assets".

5 Conclusions

In this thesis, the economic relevance of bankruptcy has been demonstrated. The impact on the firm and on the market is remarkable, and even though the event is quite rare and usually connected with overall poor performance of the company, there might also be some external events (like financial crises) which can quickly deteriorate the financial position of a firm. This is why we need effective models that can deliver accurate and reliable predictions. From the analysis performed, the use of statistical models through machine learning techniques proved to be an effective way of predicting bankruptcy. The power of these models allows the use of large datasets containing huge amounts of information, which are crucial to deliver good predictions. However, depending on the distribution of the data, some models may be more appropriate than others, and the wide selection of algorithms offers a number of possibilities to obtain good predictions. In this study, the results show that the random forest offered the best predictions of the two classes and reached satisfying results in all the performance metrics considered. Following a rebalancing of the data through SMOTE, the model achieved a 0.85 F1 score and 0.98 ROC-AUC score. The feature importance plot and the point biserial correlation highlighted that liabilities indices are positively correlated with the response, signalling that firms with high levels of financial leverage are more prone to bankruptcy. On the other hand, measures considering net income, total assets or operating efficiency (captured by variables like ROA) are negatively correlated with the qualitative response, so firms should aim at improving these values to mitigate their risk of bankruptcy. It is not a case that this model performed well on the data since the algorithm is built to be robust to outliers and is optimal when handling large datasets like this one. There are several studies that prove the random forest predictive power for bankruptcy prediction. In a similar study conducted in 2017 by Barboza, Kimura and Altman[4], the classifier proved to be one of the best models for bankruptcy prediction. Li and Wang (2018)[12] again proved the algorithm's superiority by comparing it with others. Lombardo et al. (2022) demonstrated its versatility, which is preferred to Artificial Neural Networks when taking into account computation time and costs. Not only that, the model has proved to perform better in "high dynamic contexts".

Bibliography

[1] Alam, S. I. (2022). James ohlson o-score for predicting corporate bankruptcy.

[2] Alberto Fernández, Salvador García, M. G. R. C. P. B. K. F. H. (2018). Learning from imbalanced data sets.

[3] Azzalini, A. and Scarpa, B. (2012). *Data analysis and data mining: An introduction*. OUP USA.

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

[5] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29.

[6] Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011). SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813.

[7] Chawla, N. V., Cieslak, D. A., Hall, L. O., and Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, 17(2):225–252.

[8] Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36.

[9] García, V., Sánchez, J., and Mollineda, R. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21. Special Issue on New Trends in Data Mining.

[10] Grice, J. S. and Dugan, M. T. (2001). The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher. *Review of Quantitative Finance and Accounting*, 17(2):151–166.

[11] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[12] Li, Y. and Wang, Y. (2018). Machine learning methods of bankruptcy prediction using accounting ratios. *Open Journal of Business and Management*, 06:1–20.

[13] Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research*, 252(2):561–572.

[14] Lombardo, G., Pellegrino, M., Adosoglou, G., Cagnoni, S., Pardalos, P. M., and Poggi, A. (2022). Machine learning for bankruptcy prediction in the american stock market: Dataset and benchmarks. *Future Internet*, 14(8).

[15] Luengo, J., Fernández, A., García, S., and Herrera, F. (2011). Addressing data complexity for imbalanced data sets: Analysis of smote-based oversampling and evolutionary undersampling. *Soft Comput.*, 15:1909–1936.

[16] Shi, Y. and Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital*, 15(2):114–127.

Sitography

Accountinginside, https://accountinginside.com/retained-earnings-to-total-asset/#::text=Retained%20Earning%20to%20Total%20Asset,than%20paying%20dividends%20or%20draw.

Analytics India Magazine, https://analyticsindiamag.com/curse-of-dimensionality-and-what-beginners-should-do-to-overcome-it/

App4Finance, https://www.appforfinance.com/working-capital-to-assets-ratio.html#::text=The%20working%20capital%20to%20total,the%20short%2Dterm%20company's%20solvency.

Bdc, https://www.bdc.ca/en/articles-tools/entrepreneur-toolkit/templates-business-guides/glossary/financial-statements#::text=Financial%20statements%20are%20a%20set,it%20has%20made%20and%20spent.

Builtin, https://builtin.com/data-science/skewed-data

Chartio, https://chartio.com/learn/charts/histogram-complete-guide/

Cheatography, https://cheatography.com/deleted-2754/cheat-sheets/the-altman-z-score-formula/

Classeval, https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/#::text=A%20ROC%20curve%20of%20a%20perfect%20classifier,with%20the%20perfect%20performance%20level.

Classeval, https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/#::text=A%20ROC%20curve%20of%20a%20perfect%20classifier,with%20the%20perfect%20performance%20level.

CNN, https://money.cnn.com/2006/03/24/pf/personal_bankruptcies/index.htm

Cornerstone Research, https://www.cornerstone.com/wp-content/uploads/2021/12/Trends-in-Large-Corporate-Bankrup

tcy-and-Financial-Distress-Midyear-2021-Update.pdf#::text=Bankruptcy%20filings%20in%202020%20were,in%202008%20and%202009%2C%20respectively.

Corporate Finance Institute, https://corporatefinanceinstitute.com/resources/knowledge/credit/altmans-z-score-model/

Databricks, https://www.databricks.com/it/glossary/machine-learning-models#::text=A%20machine%20learning%20model%20is,sentences%20or%20combinations%20of%20words.

Devopedia, https://devopedia.org/logistic-regression

FindLaw, https://www.findlaw.com/bankruptcy/what-is-bankruptcy/pros-and-cons-of-declaring-bankruptcy.html

Fisco e tassehttps://www.fiscoetasse.com/files/5443/modelli-previsione-insolvenze.pdf

Forbes, https://www.forbes.com/sites/steveschaefer/2011/08/10/the-great-recessions-biggest-bankruptcies-where-are-they-now/?sh=65b3088a4b7e

Geeks for Geeks, https://www.geeksforgeeks.org/best-python-libraries-for-machine-learning/

GlobeNewswire, https://www.globenewswire.com/news-release/2021/01/05/2153670/0/en/2020-Bankruptcy-Filings-Lowest-in-35-years.html

Google Machine Learning Education, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Google Machine Learning Education, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

IBM, https://www.ibm.com/cloud/learn/bagging

IBM, https://www.ibm.com/cloud/learn/random-forest

IBM, https://www.ibm.com/topics/logistic-regression#::text=Logistic%20regression%20estimates%20the%20probability,bounded%20between%200%20and%201.

IBM, https://www.ibm.com/cloud/learn/bagging

IBM, https://www.ibm.com/cloud/learn/random-forest

Imbalanced-learn, https://imbalanced-learn.org/stable/

Imbalanced-learn, https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

Intangible Capital, https://www.intangiblecapital.org/index.php/ic/article/view/1354/756

Investopedia, https://www.investopedia.com/terms/b/bankruptcy.asp

Investopedia, https://www.investopedia.com/terms/d/debtrestructuring.asp

Investopedia, https://www.investopedia.com/articles/01/120501.asp#::text=Key%20Takeaways,operate%20under%20a%20reorganization%20plan.

Investopedia, https://www.investopedia.com/articles/active-trading/081315/financial-ratios-spot-companies-headed-bankruptcy.asp

Investopedia, https://www.investopedia.com/articles/financial-theory/10/spotting-companies-in-financial-distress.asp

Investopedia, https://www.investopedia.com/terms/s/stakeholder.asp#::text=A%20stakeholder%20has%20a%20vested,%2C%20governments%2C%20or%20trade%20associations.

Investopedia, https://www.investopedia.com/terms/a/altman.asp

Investopedia, https://www.investopedia.com/terms/a/assetturnover.asp#::text=Key%20Takeaways,the%20same%20sector%20or%20group.

Investopedia, https://www.investopedia.com/terms/r/return_on_total_assets.asp

Javatpoint, https://www.javatpoint.com/logistic-regression-in-machine-learning

Machine Learning Mastery, https://machinelearningmastery.com/imbalanced-classification-is-hard/#::text=It%20is%20a%20problem%20typically,examples%20from%20the%20minority%20class.

Machine Learning Mastery, https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/#::text=Feature%20selection%20is%20the%20process,the%20performance%20of%20the%20model.

Machine Learning Mastery, https://machinelearningmastery.com/what-is-imbalanced-classification/

Machine Learning Mastery, https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Machine Learning Mastery, https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/#::text=The%20train%2Dtest%20split%20is,dividing%20it%20into%20two%20subsets.

Machine Learning Mastery, https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

Medium, https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Medium, https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

MonkeyLearn, https://monkeylearn.com/blog/what-is-a-classifier/

Neptune, https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization#::text=The%20differences%20between%20L1%20and,of%20squares%20of%20the%20weights.

Neptune, https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization#::text=The%20differences%20between%20L1%20and,of%20squares%20of%20the%20weights.

Neptune, https://neptune.ai/blog/fighting-overfitting-w

ith-l1-or-l2-regularization#::text=The%20differences%20between%20L1%20and,of%20squares%20of%20the%20weights.

Neptune, https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization#::text=The%20differences%20between%20L1%20and,of%20squares%20of%20the%20weights.

Oclc, https://www.oclc.org/content/dam/oclc/reports/2010perceptions/thegreatrecession.pdf

PayPlan, https://www.payplan.com/debt-solutions/bankruptcy/fraud-explained/

ProjectPro, https://www.projectpro.io/recipes/find-optimal-parameters-using-randomizedsearchcv-for-regression

Real Business Rescue, https://www.realbusinessrescue.co.uk/liquidation/the-difference-between-liquidation-and-administration#::text=The%20primary%20difference%20between%20the,before%20dissolving%20the%20company%20completely.

RHM firm, https://www.rhmfirm.com/blog/2017/december/bankruptcy-around-the-world/

Scikit-learn, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

Scientific Research Publishing, https://www.scirp.org/journal/paperinformation.aspx?paperid=80455#return29

Seaborn, https://seaborn.pydata.org/introduction.html

Seldon, https://www.seldon.io/decision-trees-in-machine-learning#::text=Decision%20trees%20are%20used%20as,features%2C%20a%20process%20called%20pruning.

Seldon, https://www.seldon.io/decision-trees-in-machine-learning#::text=Decision%20trees%20are%20used%20as,features%2C%20a%20process%20called%20pruning.

Chron, https://smallbusiness.chron.com/effects-rising-rates-balance-sheet-70261.html

Statistics How To, https://www.statisticshowto.com/lasso-regression/

Stockopedia, https://www.stockopedia.com/ratios/market-value-of-equitybook-value-of-total-liabilities-5041/#::text=Stockopedia%20explains%20Market%20Value%20of,and%20the%20firm%20becomes%20insolvent.

Towards Data Science, https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Towards Data Science, https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Towards Data Science, https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

Towards Data Science, https://towardsdatascience.com/what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e

TradingView, https://it.tradingview.com/support/solutions/43000597850/

Unbiased, https://www.unbiased.co.uk/life/small-business/bankruptcy-vs-insolvency-what-s-the-difference#::text=Bankruptcy%20is%20a%20legal%20process%20or%20court%20order%2C%20while%20insolvency,sole%20traders%20with%20unlimited%20liability.

University of Twente, https://essay.utwente.nl/65464/1/Kleinert_MA_Management%20and%20Governance.pdf

Varsity Tutors, https://www.varsitytutors.com/hotmath/hotmath_help/topics/normal-distribution-of-data

Wikipedia, https://en.wikipedia.org/wiki/Bankruptcy

Wikipedia, https://en.wikipedia.org/wiki/Financial_ratio

Wikipedia, https://en.wikipedia.org/wiki/Dot-com_bubble

Wikipedia, https://en.wikipedia.org/wiki/Financial_crisis_of_2007%E2%80%932008

Wikipedia, https://en.wikipedia.org/wiki/COVID-19_pandemic

Wikipedia, https://en.wikipedia.org/wiki/Bankruptcy_prediction

Wikipedia, https://en.wikipedia.org/wiki/Ohlson_O-score

Wikipedia, https://en.wikipedia.org/wiki/Machine_learning

Wikipedia, https://en.wikipedia.org/wiki/Big_data

Wikipedia, https://en.wikipedia.org/wiki/Statistical_classification

Wikipedia, https://en.wikipedia.org/wiki/Python_(programming_language)

Wikipedia, https://en.wikipedia.org/wiki/Confusion_matrix

Wikipedia, https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Wikipedia, https://en.wikipedia.org/wiki/Statistical_classification#Binary_and_multiclass_classification

Wikipedia, https://en.wikipedia.org/wiki/Binary_classification

Wikipedia, https://en.wikipedia.org/wiki/Feature_selection

Wikipedia, https://en.wikipedia.org/wiki/Oversampling_and_undersampling_in_data_analysis

Wikipedia, https://en.wikipedia.org/wiki/Confusion_matrix

Wikipedia, https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Wikipedia, https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

WilsonField, https://www.wilsonfield.co.uk/closing-limited-company/company-bankruptcy/#::text=Company%20bankruptcy%20is%20a%20term,the%20equivalent%20of%20company%20bankruptcy.