Credit Rating Estimation In Case of Missing Balance Sheet Information for Small German Companies

Enrico Regolin, MBA XXX

August 14, 2020

Contents

Chapter 1

Introduction

A key for the success of any business, is the ability to correctly assess how creditworthy business partners are. This is valid for any entity: investors, suppliers, costumers, financial institutions, etc. Typically, creditworthiness is measured in terms of credit rating, with special institutions assigned to this task. The system, however, is designed in such a way that companies looking to issue debt are charged with the credit rating costs, so that only large corporations, typically listed, can use the services of the well known Credit Agencies (Moody's, S&P, Fitch).

The large part of the economic landscape, however, is composed of small enterprises, which nevertheless require access to credit for their business activities [8]. This has generated the necessity for alternative ways of calculating a credit score for a large number of entities, with limited data available, due to looser regulation in their publication duties. From this need, models were originated, generally referred to as "Probability of Default" models (PDs) [35], which could be applied to large sets of data almost automatically.

In order to successfully generate and exploit PDs, large amounts of historical data are needed, including fundamentals such as financial ratios and all default events [42]. Moreover, this data typically is clustered per industry in which companies operate, since dynamics tend to vary considerably from one sector to another. Typically, there are specialized firms in extrapolating this kind of data and selling it to those interested in it (in general stakeholders of a particular company). However, in some cases, large companies tend to develop in house their PDs.

The data used for PDs is derived from the financial statements which companies have to publish: from these, financial ratios are computed describing size, financial leverage, liquidity, profitability, and various efficiency measures. Such ratios are recognized as driving factors to determine the likelihood for a company to default. By complementing this information with the history of the companies which defaulted, statistical regressions can be computed, used to determine which companies are likely to default in the future. These regression models provide a response in terms of probability, which can then be converted into the credit rating scales used by the agencies [8].

Like all data-driven models, PD can only be as reliable as the quality of the historical series used to generate them. Theoretically firms should only be compared within the same industry, and not over a time span which is too long (due to changing market dynamics), however by doing so the information pool would shrink, making the extrapolation of a model difficult, if not impossible. Data quality is in general an issue, since available financial records contain wrong entries as well as several missing fields, in most cases due to different regulations in certain countries, which allow companies to not disclose parts of their books. Moreover, even the default information can present time offsets or not be included at all, due to prolonged lawsuits or acquisitions [42].

Here we analyze the problem of missing data, which analysts face when trying to apply PDs to small companies: several methods can be used to ”fill” the gaps, although each one of these tend to introduce noise into the dataset. We will compare two different methods, which exploit recent statistical and Machine Learning techniques. The first one, based on an Artificial Neural Network, works well with raw datasets (in our case balance sheets) with missing data, and is a ”black-box” approach. The second approach is a more analytic/statistical one, in which financial ratios are first derived, and then a statistical explanation is used to fill the gaps. In this second try, a ”simpler” but reliable technique, known as Random Forest, is used to predict the credit rating of the companies.

The purpose of this double approach (black box vs. analytic) is to analyze the nature of the data: we perform a preliminary analysis, in order to provide a possible direction to follow in order to identify how the missing data should be handled. A deep study on this subject could be beneficial to both regulators and credit rating Agencies, to understand the patterns which cause certain companies in specific industries to omit certain accounts in their published books, and of course the analysts to deal with this aspect.

The considered dataset, provided by ModeFinance, comprises three years of balance sheet accounts of 10.000 German firms. ModeFinance is a young company, which was started in Trieste as a University spin-off operating in the FinTech sector. It exploits Artificial Intelligence technologies (in particular Machine Learning) to compute the credit rating of companies. Starting from year 2015 it is also listed by the ESMA (European Securities and Markets Authority) [4], as a licensed Rating Agency for firms and financial institutions [11]. Modefinance develops application for professionals and small-medium-enterprises (SMEs), which allows them to access financial information of all firms in Europe. Their services exploit the latest AI and big data technologies, combined with financial insight, in order to simplify their costumers' background research when dealing with clients, suppliers and financing sources.

This work is structured as follows. In Chapter 2 we investigate the legislative reasons which generates the gaps in the available information, in particular for German SMEs. In Chapter 3 we analyze the financial ratios which can be used as drivers to predict a company's credit rating, when only the balance sheet information is available. Chapter 4 presents an analysis of the datasets, based on statistical and ML techniques. Finally, in 5 we draw some conclusions*.

*The source-code of all results published in Chapter 4 is available in the GitHub repository `https://github.com/EnricoReg/CreditRatingAnalysis---share/blob/master/main_notebook.ipynb`

Chapter 2

Financial Statements Publication for SME in Germany

2.1 Financial Statements Publication and Accounting Standards

International Financial Reporting Standards (IFRS) set basic guidelines with the goal that budget reports can be reliable, straightforward and comparable worldwide. The rules in the IFRS are identified by the International Accounting Standards Board (IASB), which determine how organizations must hold their records, and report any financial transaction which occurred [6].

The basic idea behind IFRS is to have a common accounting language for all companies, with the goal that organizations and their budget reports can be predictable and dependable from organization to organization and nation to nation. Organizations profit by the IFRS in light of the fact that financial investors are more likely to finance a company if the organization's strategic policies are straightforward [7].

IFRS started in the European Union, with the goal of making business undertakings and records available over the mainland. The thought immediately spread outside the European Union, since a shared language permitted more noteworthy correspondence around the world.

To this point, the objective to make universal examinations as simple as conceivable hasn't completely been accomplished , on the grounds that, notwithstanding the U.S. utilizing Generally Accepted Accounting Principles (GAAP), a few nations utilize different norms. Synchronizing accounting principles is therefore an ongoing effort in the worldwide bookkeeping network [7].

The compliance with IFRS for the following aspects of business practice has been deemed mandatory:

- Statement of Financial Position, also known as a Balance Sheet, where the IFRS standard prescribes how the accounts should be included;

- Statement of Comprehensive Income, which can be a unique document or can comprise (a) statement of profit and loss statement and (b) statement of other income, including property and equipment for the given period;

- Statement of Changes in Equity, summarizing the company's records of earnings and profit;
- Statement of Cash Flow, with the company's monetary transactions over the financial period, detailing the cash flows as "Operating", "Investing", and "Financing".

These documents are normally shown together with the reports of one or more previous financial periods, in order to facilitate a comparative analysis and show trends. An organization should as well provide a report of its accounting policies together with the aforementioned documents. Finally, separate reports for the subsidiaries should be published by a parent company.

2.1.1 Differences between IFRS and German GAAP

Starting from 2005, European Union regulation obliged publicly traded European companies to publish their consolidated financial statements in compliance with IFRS. In practice, many German companies had already migrated from the German statutory accounting and reporting requirements (German GAAP, or alternatively Handelsgesetzbuch, HGB) and started compiling their reports according to those standards for over a decade, in order to facilitate their access international capital funding [34].

However, in the case of German non-listed and Small-Medium-Enterprises (SME), this choice is still an optionality. Reasons which are pushing companies towards one standard or the other, have been recently investigated in [17]. This study shows how private firms opting for the IFRS standard tend to grow more, use more financial leverage and are prone to issuing bonds for self-financing, hire external auditors and rating agencies, and in general have more internationally integrated business models.

Commercial law in Germany, to this date, still requires the application of German GAAP, in particular for tax purposes, distribution of profits, statutory presentation and disclosure [25, 37]. This makes the adoption of the IFRS standard in separate entity financial statements non-mandatory, and therefore it's only applied for special presentation needs, such as the access to external financing sources.

In the next paragraphs, we will highlight the main differences between German GAAP and IFRS for what concerns the publication of financial statements, with special focus on the Statement of Financial Position. The reason for the choice of focusing only on the Balance Sheet, lays in the fact that for Small Enterprises this is the only document which managers of the companies are required to present, as will be discussed more in depth in 2.2. Moreover, since the scope of this work is towards the extraction of significant financial ratios for the estimation of the credit rating of companies, we will highlight which of the differences can have an impact in this regard.

In general, a trend can be extrapolated from the information which is given in the next sections, showing how the German economic environment presents less regulations and publication duties than the European counterparts.

Components of financial statements

The IFRS standard requires the publication of the documents listed in 2.1, namely* [37]:

- statement of financial position as at the end of the period;
- statement of profit or loss and other comprehensive income for the period;
- statement of changes in equity for the period;
- statement of cash flows for the period;
- notes, comprising significant accounting policies and other explanatory information;
- comparative information in respect of the preceding period.

The same holds in case of the German GAAP, for the financial statements of single-entities which are part of public companies. However, in the case of financial statements for single-entity, Cash-Flows and Changes in Equity statements are not mandatory, while a management report has to be included in any case. Moreover, SMEs have the possibility to present condensed financial statements, while not attaching reports and notes, if the details are included within the financial statements.

Finally, a major difference between German GAAP and IFRS concerns Segment Reporting, which is only compulsory for listed companies. Segment reporting concerns the operating segments of an enterprise, which have to be considered separately in the attachments to the financial statements. It is meant to provide insight to the investors and debt holders with regards to the financial results of the different operating units [1]. In German GAAP Segment Reporting is not compulsory for parent companies that already prepare the consolidated financial statements. On the contrary, public companies which are not required to deposit the consolidated financial statements, can add it.

Statement of financial position

Some of the differences between the two standards can be viewed as formal, with moderate impact on the financial ratios used for credit rating purposes. These aspects from the German GAAP side include:

- accounts are presented in increasing order of liquidity on balance sheet;
- entities with specific legal forms are required to use a separate balance sheet format, while other requirements are present for companies providing financial or insurance services;
- the practice of offsetting assets/liabilities in IFRS is only allowed under restrictive conditions, whereas in German GAAP it can even be compulsory in certain cases;

*Note that an organization may adopt different naming for the statements from those presented here.

- in IFRS non-controlling interests are presented separately within equity, while in German GAAP they are part of equity.

Other differences between IFRS and German GAAP in terms of Balance Sheet can have an impact over the availability of required information to derive the companies credit rating (as will be discussed in Chapter 3). The main issue is connected to the fact that, while the IFRS standard requires that certain accounts should be presented separately, this is not necessarily the case in German GAAP.

This is significant in the case of the current/non-current distinction is required: in IFRS current assets and liabilities are expected to be realized in the next financial period (yearly basis). German GAAP allows for a more generic allocation of current assets, which are intended to be those not for ”long-term use” in the business [37]. One consequent is that in IFRS balance sheet deferred taxes are listed as non-current, with a current/non-current break up discussed in the notes, while in German GAAP there is not a pre-determined position.

2.2 Balance sheet Publication Rules for SME

The European Commission (Recommendation of 6 May 2003) considers Small and Medium Enterprises (SMEs) those companies which satisfy the following conditions [5]:

- people employed are less than 250;
- annual turnover is less than 50M €;
- total assets are less than 43M €.

The reasons for this definition are mainly due to the necessity of assessing which companies are entitled to apply for certain funding programmes, promoted by the EU to enhance SME businesses, and also to define which ones are subject to certain rules which in general are designed to favor competition.

In the EU the publication rules for financial statements are defined at national level for non-listed companies, while the IFRS is compulsory only for the publicly traded ones. Moreover, the transparency duties tend to vary between small and medium enterprises, so that the thresholds to be considered for these distinctions have to be those defined locally. For this reason, in the following paragraph we will discuss the regulation in Germany for SMEs.

2.2.1 Financial Statements Publication Requirements (German GAAP)

In Germany all companies are required by the Commercial Code to hold accounting records, except for very small businesses [37]. The bookkeeping shall be held in German language, and include all financial and economic transactions, as well as the necessary attached documentation. The books have to comply to the GAAP standard explained in Section 2.1.

The accounting books shall also comply to the German tax law, which includes specific provisions, such as the ”Grundsätze zur ordnungsmäßigen Führung und

Aufbewahrung von Büchern, Aufzeichnungen und Unterlagen in elektronischer Form sowie zum Datenzugriff" ("principles of orderly accounting and retention of books, records and documents in electronic form and of data access", see [37]). Other formal requirements pose additional constraints, such as the location where the records should be kept, which in general cannot be held outside Germany. Only under certain conditions electronic books can be made available in another countries. Within the scope of this work, we can state that these restrictions in general do not facilitate the retrieval of more in depth information required for the classification within the credit rating classes.

Another point of interest is that the Commercial Code includes guidelines for the publication of financial statements and auditing of all limited companies (GmbH in Germany). This regulation affects a large portion of the SMEs which are included in the available database, which are therefore subject to more regulation than the case in which there is unlimited liability.

	Small company	Medium-sized company	Large company
Annual sales (€)	$\leq 12M$	$\leq 40M$	$> 40M$
Balance sheet total (€)	$\leq 6M$	$\leq 20M$	$> 20M$
Employees (€)	≤ 50	≤ 250	> 250

Table 2.1: Companies classification in the EU.

In general, with regards to audit and publication of financial statements, businesses are clustered according to size, as illustrated in Table 2.2.1. The belonging of a company in one of the three clusters depends on the thresholds which, similarly to the EU classification, regard number of employees, revenues, and balance sheet. It should be noted that for an enterprise to change status, two out of three of the conditions have to be met in consecutive years.

Generally, companies must present the required documents within one year of the end of the financial period. Large businesses, and more in general public ones, or companies which issue public bonds in the German market, have a closer deadline to publish their books. Audits are not mandatory for small businesses: they are only required to present electronically a condensed statement of financial position, without income statement. On the other hand, Medium and Large-sized enterprises are required to have their books reviewed by third party auditors. Nevertheless, the publication requirements are less burdensome for medium-size companies, compared to the large ones.

In Fig. 2.2.1 the distribution of companies size (considering Total Assets) is shown for the available database. In this representation, only the presumptive SMEs (according to the EU definition) are included, which therefore are not obliged to comply to IRFS, accounting for $\approx 98\%$ of all available ones. Out of these businesses, the vast majority have Total Assets $< 6MEuro$, which makes them likely to be Small business. Finally, out of all these, the unlimited liability companies benefit from, very limited publication requirements, which explains the scarcity of information for these enterprises.

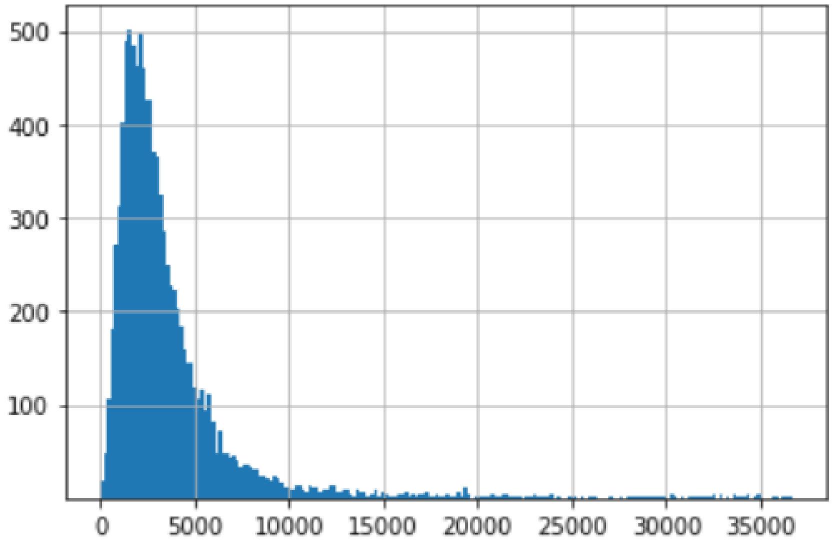

Figure 2.1: Distribution of companies in the database based on Total Assets (thousends of Euros).

2.3 Classification of companies based on the Industry

The information included in the financial statements alone often is not enough to explain the credit rating of a company. In several cases, the level of risk associated with the investment in a given business is related to the industry in which the enterprise is operating. A thorough analysis of this aspect would require to retrieve the information regarding the β of different industries, in order to understand how sensitive to market oscillations the segment in which a company operates is. The relationship between CAPM, industries β and credit rating is well documented in several works (see e.g. [31]).

A different (and simpler approach) could be to assume that the dynamics which influence the credit rating are similar within an industry, and analyze separately the relationships between the available data for each of these ones. One possible problem with this solution could be related to the considerable differences in frequency between companies operating in a segment, compared to others. A solution to this problem is to deal with the whole dataset as a unique database, assigning certain labels to to the companies operating in the same industries. In this way, by working with Data Analysis tools for classification, it can be possible to automatically provide an additional criterium to improve the qualification, without decomposing the available dataset.

Eurostat, the Statistical Office of the European Union, provides a tool for this kind of operation, with the NACE Code classification [24]. Eurostat collects statistics from the different national statistical offices in Europe, in order to provide a unique source of data which drives the choices for the policies run by the European Commission and the European Parliament. In order to do so, several standards have been enforced, with the aim of harmonizing the data available from the EU members. One of these standards is the so-called NACE, which stands for ”Nomenclature statistique des activités économiques dans la Communauté européenne” (”statistical classification of economic activities in the European Community”).

With technological advancements and radical changes in economy and society, the NACE standard has to be periodically updated, in order to be able to provide a useful statistical representation of the economy within the EU. The current version, NACE Rev. 2, has been released in 2008, after an 8-years long revision work. This latest version is now an integrated system, which is able to produce comparable statistics also in different segments (e.g. production of goods vs. trade). The entire statistical classification method is based on the following three principles:

- there should be a logical consistency in the methods used for the assignment of certain elements to a given category;
- business activities and entities operating within the EU should be universally covered;
- the identify categories should be mutually exclusive.

In the following paragraph we go more in detail with the explanation of how the NACE standard is organized, and finally present the overall table 2.2 used for the allocation of the companies.

2.3.1 NACE Classification Code

The NACE Regulation defines a hierarchical form for the NACE code. It consists of 4 levels, described as follows:

1. **Section** (alphabetical symbol);
2. **Division** (2 digits number);
3. **Group** (3 digits number);
4. **Class** (4 digits number).

The top level (**Section**) depends on the broad type of good or service which is sold by the company. At lower level, **Groups** and **Divisions** depend on the character of the goods/services produced and their final destination of use, including inputs and production process. In the final category (**Class**) the production process or the technology used to provide the service are considered.

One key aspect in this classification procedure, is the instance in which large parts of the activities of an entity should be included in more than one NACE category. Typically, these situations occurr in case of vertical or horizontal integration of activities. In these situations the business is assigned considering a ”decision tree” in which at each node, the branch with highest (relative percentage) of added value is considered.

This logic can be understood by looking at the diagram in Fig. 2.3.1. In the case a business has activities in two different NACE categories, the one with higher

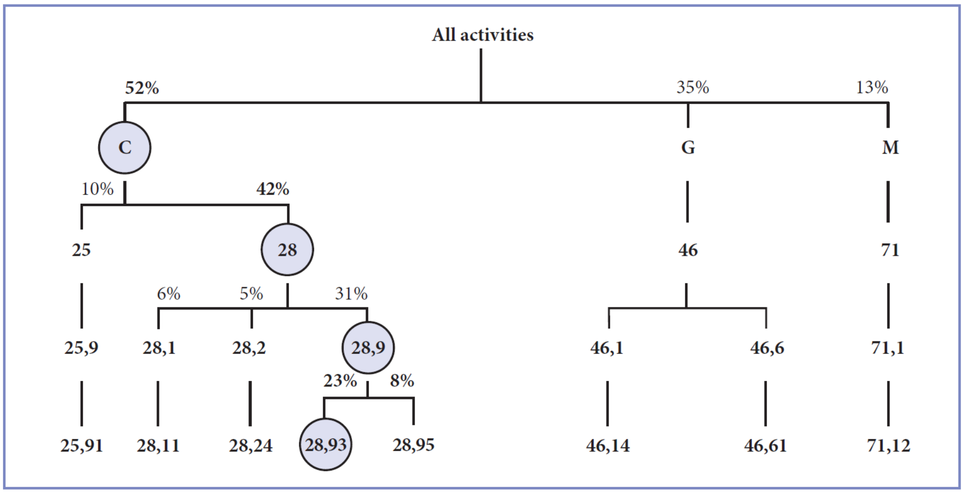

Figure 2.2: Representation of the "top-down" method for the selection of the proper NACE Class in case of entity with multiple businesses. (from [24]).

value added is considered for classification purposes in NACE Rev. 2. If the three or more activities corresponding to different NACE classes are identified, the "top-down" process is used for classification.

According to the "top-down" principle, the allocation of an entity at **Class** level should be consistent with its classification in the higher levels. Therefore, the method starts with selection of the **Section**, and continues with the lower levels, always selecting the Dsivision/Group/Class with the highest share of value added among the different options.

NACE Codes Table

In the table below, the first two levels of the NACE classification are listed. The describtions for the different Divisions are grouped for each Section, in order to provide an understanding of how the different activites are grouped.

Section	Group	Description
A	1	Crop and animal production, hunting and related service activities
	2	Forestry and logging
	3	Fishing and aquaculture
B	5	Mining of coal and lignite
	6	Extraction of crude petroleum and natural gas
	7	Mining of metal ores
	8	Other mining and quarrying
	9	Mining support service activities
C	10	Manufacture of food products
	11	Manufacture of beverages
	12	Manufacture of tobacco products
	13	Manufacture of textiles
	14	Manufacture of wearing apparel

	15	Manufacture of leather and related products
	16	Manufacture of wood and of products of wood, straw and cork, except furniture
	17	Manufacture of paper and paper products
	18	Printing and reproduction of recorded media
	19	Manufacture of coke and refined petroleum products
	20	Manufacture of chemicals and chemical products
	21	Manufacture of basic pharmaceutical products and preparations
	22	Manufacture of rubber and plastic products
	23	Manufacture of other non-metallic mineral products
	24	Manufacture of basic metals
	25	Manufacture of fabricated metal products, except machinery and equipment
	26	Manufacture of computer, electronic and optical products
	27	Manufacture of electrical equipment
	28	Manufacture of machinery and equipment n.e.c.
	29	Manufacture of motor vehicles, trailers and semi-trailers
	30	Manufacture of other transport equipment
	31	Manufacture of furniture
	32	Other manufacturing
	33	Repair and installation of machinery and equipment
D	35	Electricity, gas, steam and air conditioning supply
E	36	Water collection, treatment and supply
	37	Sewerage
	38	Waste collection, treatment and disposal activities; materials recovery
	39	Remediation activities and other waste management services
F	41	Construction of buildings
	42	Civil engineering
	43	Specialised construction activities
G	45	Wholesale and retail trade and repair of motor vehicles and motorcycles
	46	Wholesale trade, except of motor vehicles and motorcycles
	47	Retail trade, except of motor vehicles and motorcycles
H	49	Land transport and transport via pipelines
	50	Water transport
	51	Air transport
	52	Warehousing and support activities for transportation
	53	Postal and courier activities
I	55	Accommodation
	56	Food and beverage service activities
J	58	Publishing activities
	59	Motion picture, video and television programme production, sound recording and music publishing activities

	60	Programming and broadcasting activities
	61	Telecommunications
	62	Computer programming, consultancy and related activities
	63	Information service activities
K	64	Financial service activities, except insurance and pension funding
	65	Insurance, reinsurance and pension funding
	66	Activities auxiliary to financial services and insurance activities
L	68	Real estate activities
M	69	Legal and accounting activities
	70	Activities of head offices; management consultancy activities
	71	Architectural and engineering activities; technical testing and analysis
	72	Scientific research and development
	73	Advertising and market research
	74	Other professional, scientific and technical activities
	75	Veterinary activities
N	77	Rental and leasing activities
	78	Employment activities
	79	Travel agency, tour operator and other reservation service and related activities
	80	Security and investigation activities
	81	Services to buildings and landscape activities
	82	Office administrative and business support activities
O	84	Public administration and defence; compulsory social security
P	85	Education
Q	86	Human health activities
	87	Residential care activities
	88	Social work activities without accommodation
R	90	Creative, arts and entertainment activities
	91	Libraries, archives, museums and other cultural activities
	92	Gambling and betting activities
	93	Sports activities and amusement and recreation activities
S	94	Activities of membership organisations
	95	Repair of computers and personal and household goods
	96	Other personal service activities
T	97	Activities of households as employers of domestic personnel
	98	Undifferentiated goods- and services-producing activities of private households for own use
U	99	Activities of extraterritorial organisations and bodies

Chapter 3

Credit Rating and Financial Ratios

3.1 The Credit Rating System

Credit Rating is the process of quantifying numerically the trustworthiness of a borrower with regards to its existing financial obligations, or portions of it. Therefore, a credit rating can be applied to any individual, or organization which asks for a loan or issues any form of debt [13].

In the case of individuals, the Credit Score is very common in Anglo-Saxon countries, and is calculated by agencies such as TransUnion, Experian, which comply to the Fair Isaac (FICO) credit scoring.

Credit assessment for companies and governments is the bond credit rating, focusing on their potential capability to repay corporate or government bonds. Professional and investors make use of this information, published by the Credit Rating Agencies, when evaluating their investments. Bond credit ratings are typically performed by Standard & Poor's (S&P), Moody's, or Fitch, which are also known as the "Big Three" Rating Agencies. These Rating Agencies are paid by the governmental agency/company which has issued the debt. As we will see in Section 3.1.1, this specific dynamic has generated some controversy, due to the lack of independence inherently connected to this mechanism.

In Table 3.1 the credit scores for Moody's and S&P/Fitch are listed, with the relative description. For S&P and Fitch outlooks are expressed with +/- sign. In the database provided by ModeFinance, which will be analyzed in the next chapter, outlooks are not considered.

In general, credit rating plays a key role in the financial and economic dynamics of the modern society, since it affects significantly the possibilities of a potential borrower to be granted a loan or to issue new debt. At the same time, the interest rate of a new loan will depend on the credit rating of the entity at the moment of issuance. This aspect is at the root of the previously mentioned conflict of interest, since credit rating agencies are paid by those who can gain an advantage or be damaged by their decisions.

Moody's	**S&P, Fitch**	**Credit worthiness**
Aaa	AAA	An obligor has EXTREMELY STRONG capacity to meet its financial commitments.
Aa1 Aa2 Aa3	AA+ AA AA-	Obligor has VERY STRONG capacity to meet its financial commitments. It differs from the highest-rated obligors only to a small degree.
A1 A2 A3	A+ A A-	Obligor has STRONG capacity to meet its financial commitments but is more susceptible to the adverse effects of changes in circumstances and economic conditions than higher-rated categories.
Baa1 Baa2 Baa3	BBB+ BBB BBB-	Obligor has ADEQUATE capacity to meet its financial commitments. Adverse economic conditions or changing circumstances are more likely to weaken its capacity to meet financial commitments.
Ba1 Ba2 Ba3	BB+ BB BB-	Obligor is LESS VULNERABLE in the near term than other lower-rated obligors. However, it faces major ongoing uncertainties and exposure to adverse business, financial, or economic conditions.
B1 B2 B3	B+ B B-	Obligor is MORE VULNERABLE than 'BB', but currently has the capacity to meet its financial commitments. Adverse business, financial, or economic conditions will likely impair its capacity or willingness to meet the financial commitments.
Caa	CCC	Obligor is CURRENTLY VULNERABLE, and is dependent upon favourable business, financial, and economic conditions to meet its financial commitments.
Ca	CC	An obligor is CURRENTLY HIGHLY-VULNERABLE.
	C	The obligor is CURRENTLY HIGHLY-VULNERABLE to nonpayment. May be used where a bankruptcy petition has been filed.
C	D	An obligor has failed to pay one or more of its financial obligations (rated or unrated) when it became due.

Table 3.1: Credit scores description. Table taken from [13].

3.1.1 Independent Credit Rating Agencies

The credit rating industry is extremely concentrated at global level, as the Big Three agencies control the market. These three companies started analyzing credit ratings in the early XX century, creating a sort of monopoly in their field. However, their role in the 2008 financial crisis raised increasing concerns among those who question whether their business model grants them enough independence in their evaluation, considering that they are obtaining their profits from the same companies they are judging [40].

The issue regarding the independence of credit rating agencies is a very complex one, which deals with institutional and legislative aspects. The increasing attention to what exactly are the evaluation criteria used by credit rating agencies is justified by the impact these decisions can have, at market level, but also their level of influence on sovereign governments, which bonds yields are also influenced by these agencies.

With regards to these aspects, the 2008 financial crisis put under the collective lens the dramatic consequences of an inadequate credit rating, when the Big Three failed to properly evaluate financial products which were derived from deteriorating credits. Eventually this led to the collapse of Lehman Brothers, Bear Stearns and other investment banks, as well as the largest worldwide economic crisis in 80 years [26].

Following these events, the role of small credit rating agencies has increased in the past decade [29], in part thanks to the reduced reputation of the Big Three's. In fact, these have also been accused of having a negative "neighborhood bias", i.e of having a softer approach with American companies compared to the European counterparts. This fact has pushed many operators to entertain the idea of a European Credit Agency [16], although at the moment there have not been official steps in this direction.

3.2 Credit Rating Based on Financial Ratios

Financial ratios have been identified as one of the best tools to assess the credit rating of companies since several decades [27, 18].

More recently, the focus has shifted towards the possibility of predicting changes in the credit rating, in order gain an edge and use information before it is available to the market [22, 28].

Having in mind the available database mentioned in the previous chapter, in the following paragraphs we analyze the financial ratios suggested in [22] for this purpose, focusing on those which can be computed from the balance sheet. We divide these ratios into two groups:

- Financial ratios directly computable from balance sheet;

- Financial ratios NOT directly computable from balance sheet.

As we will see, some approximation are included in order to use also the ratios from the second group.

3.2.1 Financial Ratios Directly Computable From Balance Sheet

Current Ratio

Repaying short-term obligations is one of the most important aspects a company has to face, in order to keep the business running. The capability of doing it is captured, among other ratios, by the Current Ratio (CR), which is sometimes also referred to as "working capital" ratio. Typically this liquidity ratio concerns obligations and credits expiring within the financial period (1 year). Investors need this information, since a company with scarce liquidity is a very risky investment, where capitals may be used for short-term needs rather than long-term investments. CR is computed as follows:

$$CR = \frac{\text{CurrentAssets}}{\text{CurrentLiabilities}}$$

Current Assets and Current Liabilities are both present in any enterprise's Statement of Financial Position: the first one includes Cash & Cash Equivalents, Account Receivable and Inventory. Account Receivables and Inventory are supposed to be converted into cash within one financial year, or else they should be listed in a different account. Current Liabilities, on the other hand, considers all short term obligations, such as Account Payable, Wages and Taxes, as well as the installments of the long-term debt due within the financial year.

Similarly to most financial ratios, CR for a given company should be compared to other enterprises in the same industry: values similar to or slightly above the average are considered a positive indicator, while values below may be a symptom of financial distress. At the same time, very high values of CR are not necessarily positive things, as they may indicate that the management of the company is not allocating resources in an efficient and productive way.

Relying too much on the CR for evaluating the liquidity of a company may carry its own risks: in particular it is difficult to compare it across different industries, and this extends to large groups operating multiple and diverse activities. Moreover, similarly to all other ratios we will consider here, CR should be evaluated together with previous years information, in order to extrapolate a trend, more than a single figure.

Quick Ratio

The short term liquidity of a company is also captured by the Quick Ratio (QR), also known as "Acid Test". The Quick Ratio is used to assess a company's capability to repay its financial obligations due in the financial period, by only using liquid assets, such as Cash and Accounts Receivable, without selling its Inventory or using other sources of financing.

What differentiates QR from CR, is that only near-cash assets are considered, i.e. assets which can be theoretically converted into cash on the spot. Thus making it a more conservative metric. Therefore, in the following equation, Inventory is not considered for this computation:

$$QR = \frac{CE + MS + AR}{CL} = \frac{CA - I - PE}{CL}$$

with CE being Cash & Equivalents, MS Marketable Securities, AR Accounts Receivable, CL Current Liabilities, CA Current Assets, I Inventory and PE are Prepaid Expenses.

Similarly to the CR, the QR should be compared with the industry average, with low level associated with financial distress, and high level with high liquidity. However, the ideal Quick Ratio should yield a result $QR = 1$, which implies that the enterprise has planned its liquid resources to match exactly the short term liabilities they are supposed to pay off within the financial year. Values of QR below such threshold might worry management and investors, as they signal some potential liquidity issue.

Inventory and other current assets are not considered in the QR, since it is not always possible to immediately liquidate them, although the degree of difficulty in doing it varies considerably from business to business: in general it requires some time to convert Inventory into cash. Finally, another omission are Prepaid Expenses, since they cannot be used to pay off any financial obligation.

Cash Ratio

The final metric which measures the liquidity of an enterprise is the Cash Ratio (CashR), i.e. which is an even more conservative ratio than QR, since it only considers Cash and Cash Equivalents at the numerator. These include money market accounts funds, savings accounts, and treasury bills, so that they are sometimes also referred to as marketable securities.

This metric is very similar in nature to CR and QR, and one could look at it as a worst-case scenario for the CR, in which Inventory and Accounts Receivable lose all their value, as if the the company were on the verge of shutting down, so its formula is (using the acronyms defined above):

$$CashR = \frac{CE}{CL}$$

In a bankruptcy situation, the Cash Ratio provides to the creditors what percentage of the company's current liabilities could be covered with its most liquid assets, since it is assumed that the sale of other current assets, like Inventory and Accounts Receivable, is unlikely to yield their face value.

Debt to Equity Ratio

The Debt-to-Equity Ratio (D/E) is one of the most well known financial tools, which expresses the balance between long term liabilities owned by creditors and the total shareholder's equity, thus measuring the financial leverage of the company. Its formal definition is the following:

$$D/E = \frac{\mathrm{LongTermDebt} + \mathrm{ShortTermDebt}}{\mathrm{Shareholders'Equity}} \tag{3.1}$$

The Debt-to-Equity Ratio affects the cost of capital for a firm, and additionally it indicates how much of the shareholder's equity should be used to to cover all outstanding debt, in case of a shut down of the business, or alternatively which share of the debt could be theoretically recovered from the company's equity.

The exact accounts used to build numerator and denominator in the formula may deviate from the traditional accounting definition of ”debt” or ”equity”, since there are several elements which could produce a distortion of D/E: therefore the calculation of a company’s leverage is typically not straight-forward, and has to account for several distinctions in terms of intangible assets, retained earnings/losses, and pension plan adjustments.

Moreover, there is not always an agreement among analysts in the definition of debt. An example is the Preferred Stock, which is often classified as Equity, but contains elements which make it closer to Debt (preferred dividend and liquidation rights). In certain situations, the choice of including it in one category over the other could produce a considerable swing in the financial leverage output.

This kind of low-level analysis depends on the purpose of the computation: in our case we deal with a large data-set of companies, so that an individual analysis of each company’s balance sheet could not be performed, and the output of Equation (3.1) is assumed to be correct.

In general, while higher leverage can represent a risk for both shareholders and debt holders, it is very difficult to compare the D/E ratio among enterprises operating in different industries, since capital needs change depending on the business type and the growth rate: slow growing, capital-intensive industries such as automotive sector or utilities typically show high D/E, while services and high-tech sectors have a much lower one. The reason for this is that the interest rates on debt remain quite low even with high leverage, in businesses with relatively stable income: therefore this is an efficient use of capital, as opposed to other sectors, where high interest rates make it preferable to keep a low financial leverage.

3.2.2 Financial Ratios Indirectly Computable From Balance Sheet

Not all financial ratios can be computed directly from the data included in the balance sheet. In fact, several indicators require information such as Revenues, Cost of Goods Sold, Net Income and Financial Interests paid, which are contained in the Income Statement. Most of these accounts can’t be approximated with the available information, so that useful ratios such as Interest Coverage Ratio, Inventory Turnover Ratio, Cash Conversion Cycle and Operating Profit Margin can not be exploited in this work.

However, one key information could be approximated by comparing balance sheet in consecutive years. Under the (strong) assumption that all earnings/losses are retained, with no dividends paid to shareholders, no capital increases. Such assumption, although strong, could not be made in case of listed companies, due to the impact of stock price oscillations. Nevertheless, here we focus on enterprises which don’t have any publishing duty regarding their Income Statement, and therefore by definition can’t be publicly traded.

The idea is that the difference between the Total Shareholders’ Equity over consecutive years is equal to the Net Income of the company in the first year:

$$\text{NetIncome}(0) \approx \text{TotalEquity}(1) - \text{TotalEquity}(0) \tag{3.2}$$

With the approximation in (3.2), one hase the data to obtain approximation of two important ratios, Return on Assets and Return on Equity, which we will analyze below.

Return on Assets

The purpose of a business is to extract value from its resources, i.e. its assets. Therefore, a key performance indicator for a company's efficiency is the measure of how much value is generated from its assets, hence the name Return on Assets (ROA), defined as follows:

$$ROA = \frac{NetIncome}{TotalAssets}$$

Investors obtain from ROA a quantitative measure of how efficiently the investments are turned into cash. Differently from the ratios seen before, there is theoretically no downside as ROA increases, since it just indicates that the business is organized in an effective way, whereas high CR or QR could mean that an excess of cash is hiding a suboptimal resources allocation.

Even more than all other indicators we've seen thus far, ROA can't be evaluated as a standalone figure, but it should be either compared to the value in previous years, or to the average of companies in the same industry. Financial services or capital intensive manufacturing industries require a much greater amount of assets in their productive cycles, compared to businesses active in web services. For this reason ROA can't be compared across these industries. However, when focusing on improving business efficiency, evaluating the year-over-year changes in ROA is one of the best corporate evaluation metrics.

ROA doesn't contain any information about the company's financial structure, as it doesn't differentiate between Debt and Equity. However, they are both used as financing sources in order to fund the business operations. That means that the ROA of an enterprise which relies heavily on debt, might result considerably higher than an equity-only company, even if the latter is run less efficiently. This occurs because interest expenses are deducted from earnings in the Net Income.

In order to account for this, in some cases, the cost of interest used to acquire the assets is added to the numerator included in the ROA formula, in order to give a more objective portrait of the business efficiency, which is the ultimate goal of the ROA indicator.

Return on Equity

The final indicator considered is Return on Equity (ROE). It is defined as the ratio between Net Income and total equity (typically being expressed in percentage terms):

$$ROE = \frac{NetIncome}{TotalShareholdersEquity} \tag{3.3}$$

It should be noticed that, similarly to the ROA case, such indicator makes sense if both numerator and denominator are positive in (3.3), otherwise this metric might be misleading.

For public companies, the considered Net Income should be the one before dividends are paid to common stock holders, but after preferred shareholders received their share and interest expenses have been paid to creditors. In our case, dealing with non listed small companies, this aspect does not represent an issue, and approximation (3.2) is used.

All considerations expressed for the previous ratios, in terms of limitations of this indicator, are valid also for ROE: the ratio does not contain an absolute information alone, but should be compared to previous years, and to other companies in the same industry (and with comparable financial leverage).

As a measure of business efficiency, ROE has one major drawback, which should be considered. A viable way to increase considerably ROE is to issue long term debt, in order to create a large assets base to increase the revenues of the business. This is a common way to increase the business, but at the same time it increases the risk for shareholders, due to financial leverage. From this point of view, ROE can be viewed as a less reliable way to measure operational efficiency than ROA. For large, publicly traded companies, the target ROE is the market benchmark, e.g. the S&P 500 average return of $\approx 14\%$: ROE values below 10% are considered unsatisfactory from investors' point of view.

Earnings after interest expenses and taxes concur to determine the Net Income over a financial period, while Total Shareholders Equity is the average shareholders' equity, calculated by averaging out the values at the beginning and at the end of the same period. ROE is computed from the average equity to account for the different time-lines of balance sheet (photography of the situation at the end of the year) and income statement (representation of one year worth of financial and economic transactions). Given these definitions, it appears straightforward that, from an investor's point of view, this is the best metric to be used to evaluate potential buy/sell actions on the company stock, in case of publicly traded companies.

Albeit these considerations, ROE can be a dangerous tool if used out of context. In fact, in most situations, a business with an average but steady ROE is to be preferred to a company with ROE which displays extreme spikes. The reason lies in the dynamics between numerator and denominator in (3.3): if a business is facing a negative period, it might accumulate operational losses, and retain them, reducing the total equity over time. In these conditions, a sudden increase in Net Income, possibly due to investments paid with new issued debt, might generate a huge spike in ROE, also thanks to a reduced denominator. However, from a long term investor point of view, it might have been preferable to own shares of a company with a constant positive ROE, although with ”smaller highs”.

In conclusion, one can say that ROE is one of the most interesting indicators, and at the same time one with several drawbacks. One can be reasonably sure that, if ROE is negative or has excessively high values, there should be a deeper analysis to determine the roots of this behavior. Reasons for such situations can vary considerably: buyback program could produce negative ROE even in optimally managed corporations. For all accounts, it should be avoided to use ROE to analyze businesses when they display negative values. In this circumstances other indicators should be considered instead.

Chapter 4

Data Analysis for Credit Rating with Incomplete Information

4.1 Problem Description

This work deals with the task of finding suitable solutions to the problem of identifying the credit rating of SME enterprises, in situations in which several accounts of the balance sheet are missing. In order to propose a solution, the given database of German companies can be handled in different ways.

In particular, we focus on two approaches: at first a pure black-box approach is used, based on an Artificial Neural Network (ANN). ANNs are powerful tools which are used in multiple fields, in order to explain complex phenomena. At the same time, their architecture allows to use them also in the case of databases with missing items.

The second approach is a more ”grey-box” one, and it requires the usage of proper financial ratios. In the case in which no Income Statement is available, some financial ratios have to be substituted with proxies, which are estimated from the Statement of Financial Position. In this approach, the database is split between the entries with all financial ratios available, and those with missing data. Then, a Machine Learning technique for classification is used (Random Forest), where the classifier is trained by using data from the fully available fields. An Expectation Maximization (EM) Data Imputation technique is then used to fill the missing fields from the database, based on the assumption that all entries are independent and identically distributed variables extracted from a Multivariate Normal Distribution. The success in the imputation process by means of the EM iterative algorithm is not guaranteed, mostly because there is no a-priori knowledge about the distribution of the data. Nevertheless, as it will be shown, the imputed dataset performs well on the Random Forest classifier trained with the full data.

The idea behind the comparison between these two approaches (”black-box” ANN, and ”grey-box” EM Imputation/Random Forest) is that the ANN is used as a benchmarking tool for the classifier based on financial ratios. In fact, for a static database such as the one available, a properly trained ANN with linear layers should be able to fully explain the relation of the labels with the features, based on the existing correlation (i.e. the residuals should not contain information). In other

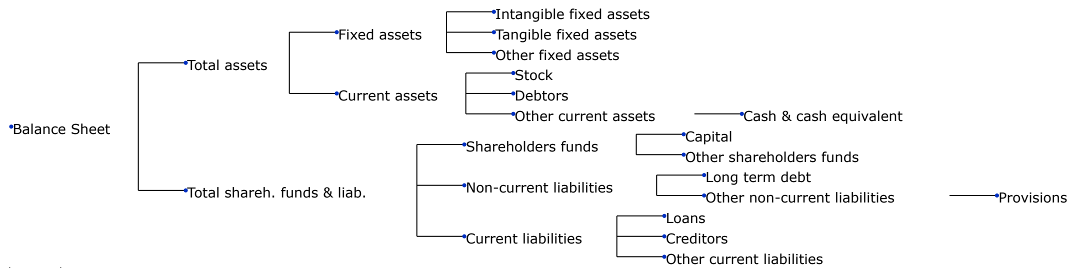

Figure 4.1: Tree diagram with the balance sheet accounts included in hierarchical levels.

words, the ANN should be able to capture the entire information which is present in the database. Therefore, if the EM/Random Forest approach has comparable performances, one can reasonably state that the choice of financial ratios, and most of all of the approximations, was justifiable.

4.1.1 The available Database

A database comprising the Statement of Financial Position of 10.000 German companies (mostly SMEs) is given, covering a three years period (2016-2018). For each year, 22 Balance Sheet accounts are available, together with the corresponding estimated credit rating classes as computed by Mode Finance. In this work we assume that the given credit rating score is the "correct" one, and we will investigate possible solutions to estimate it starting from the balance sheet information, also in the case of missing input data.

The hierarchical structure of the available accounts is shown in the picture Fig. 4.1. Four levels are identified in this structure:

- Level 0: Assets vs. Liabilities & Equity;
- Level 1: Distinction between fixed and current Assets, fixed, current Liabilities and Shareholder's Funds;
- Level 2: More detailed description regarding the nature of the accounts (e.g. Loans vs. Creditors);
- Level 3: Further details on certain accounts (e.g. Cash and Equivalents component of the account Other Current Assets).

In our analysis of the missing information, we focus on Levels 1 and 2, which provide an exhaustive and symmetric explanation of the Balance Sheet figures, not considering Provision and assuming Other Current Assets is entirely composed of Cash and Equivalents (therefore neglecting Level 3).

Some pre-processing of the original dataset was also required, as follows:

- the database is extended to include a feature with the NACE classification code (data included in the "Section" field, see information in 2.3.1) for each company. This categorical entry is one-hot encoded (see [14]) in order to train Neural Network and the Random Forest classifier;

- an additional enhanced database is created with financial ratios derived from the Balance Sheet information (see Section 3.2.1), with the aim of retaining the key financial information with reduced number of features;
- in both databases, only SME are considered, by selecting companies with Total Assets equal or less than 43M€, as per German regulation. Even with this filtering, the database size is substantially unchanged, as SME account for more than 98% of its entirety;
- each year is considered as an independent entry, so the database appears one of 30.000 separate companies, with 22 features plus the one-hot encoded NACE code.

Missing Features Statistics

Account	Missing Data [%]
Fixed assets	0.00
Current assets	0.00
Shareholders funds	0.01
Non-current liabilities	0.00
Current liabilities	0.00

Table 4.1: Percentage of Missing Data: Level 1

The database is first analyzed in terms of missing data. From Table 4.1 one can see that the ”Level 1” accounts (Fixed assets, Current assets, Shareholders funds, Non-current liabilities, Current liabilities) do not present systematic trends of missing data, since their publication is mandatory by law. Therefore we concentrate on ”Level 2” items.

Account	Missing Data [%]	Group
Intangible fixed assets	2.13	A
Tangible fixed assets	2.13	A
Other fixed assets	2.13	A
Stock	0.81	B
Debtors	0.81	B
Other current assets	0.20	C
Capital	1.92	D
Other shareholders funds	2.01	D
Long term debt	0.11	E
Other non-current liabilities	0.11	E
Loans	43.01	F
Creditors	42.94	F
Other current liabilities	37.09	F

Table 4.2: Percentage of Missing Data: Level 2

As one can understand from the diagram in Fig. 4.1 and the frequency of missing data in Table 4.2, there are 6 main ”Level 2” groups in which data may be missing,

Figure 4.2: Cross-correlation of missing data of the Level 2 Balance Sheet accounts.

which are clustered accordingly in the table (letters A,...,F). These groups are chosen since in most instances, figures missing from different accounts in the same group occurs at the same time (e.g. Stock and Debtors). The correlation between the occurrencies of missing data among these 6 categories (see Fig. 4.2) shows that, when some info is missing on the asset side (current or non-current), it is likely that also others are missing, while there is not a significant correlation between info on the asset side and on the liabilities side, and within this last group.

Based on these considerations, we identify 3 clusters of missing data, which we will then use to evaluate the performance of the classification/regression methods for credit rating classes:

- ”Asset Side” missing data (Intangible/Tangible fixed assets, Stock/Debtors, Other Current assets, Capital/Other shareholders funds);
- Long Term Debt missing data;
- Loans missing data.

This seems a reasonable choice, since an analysis between the three groups and the credit rating (Fig. 4.3) shows how there is not significant correlation among the three clusters, and between them and the credit score.

4.2 Artificial Neural Networks

The use of regression models for the description of datasets is widespread. These Regression Models, which are very useful for prediction purposes, can have varying degrees of complexity, from simple Linear and Polynomial ones, to Logistic and Least Absolute Shrinkage and Selection Operator (LASSO) [30]. However, as the complexity of the underlying behavior increases, due to high dimensionality, complex relations

Figure 4.3: Cross-correlation of the three chosen clusters of missing data with the credit rating score.

among the features or gaps in the data, the traditional Regression models could prove difficult to fit on a given dataset.

In this context, starting from the early 2000s, methods of pattern recognition based on Machine Learning, rather than traditional statistics, have gained traction in several fields, from bio-technologies to fin-tech, both in the research environment and in the industrial world [19].

In this section we will analyze how to use an Artificial Neural Network (ANN) for the purpose of identifying relationships between the different features present in the dataset, and the credit rating assigned to the companies by Mode Finance.

4.2.1 Use of ANNs in the Financial World

In the context of finance and economics, Artificial Neural Networks have mostly been employed with the purpose of predicting stock markets movements, based on large series of historical pricing data. Scientific literature is quite abundant in this field, although the stock market prediction is actually one of the most difficult tasks to accomplish in the field of time-series predictions, due to the complexity of the correlations between signals, the considerable noise and the volatility of prices.

Information is the cornerstone of any activity in this field. The automatization and digitalization which took place in the financial world, has generated extremely large and differentiated datasets which are available to operators, although at the same time they make the analysis of the information even more challenging. Moreover, the analysis is complicated also by the features present in the data, which are quite intertwined with the heavy impact of unpredictable human activity associated with these signals.

A survey of the different methodologies applied can be found in [20], where the different aspects revolving around the extraction of information from financial time-series are covered: data-mining of financial news, pre-processing and clustering of the available data, predictions on movements of the market, and many other aspects.

Another interesting work in this field is [21], in which they analyze the current state of the art in the application of several Machine Learning techniques for the analysis and prediction of financial data related to companies exchanged on the stock market. From [21] it emerges that Machine Learning's capability of extracting information from unlabeled datasets makes it particularly valuable for high frequency predictions in the stock market, in particular when using deep learning.

While the cluster of deep learning algorithms is extremely wide, in terms of structure of the ANN and various model hyper-parameters, the heaviest dependency appears to be data representation, so that pre-processing is one of the most important steps. In [21] they consider these optics, and analyze 3 different unsupervised learning methods using as inputs intra-day stock market prices, to evaluate the ANN's potential in predicting the market.

Currently, the methodologies used in this world are quite advanced, taking advantage of the latest developments in ANN theory. In [15], for instance, multiple ANN types are used in a combined framework, with stacked autoencoders (SAEs) and long-short term memory (LSTM). In [15] SAEs are at the core, with the task of learning the patterns of pricing histories (unsupervised learning), with the target of allowing SAEs to learn the invariant features of the signals.

As we have seen, the application of ANN in the field of stock market prediction is extremely developed, as the profit-driven traders have moved to a more technologically intense approach. On the other hand, other aspects of the world of financial analysis have not been exploited as much, at least in terms of publications. This is the case of the data-based credit rating estimation. In the next sub-section we will present an application of an ANN composed of linear layers, with the aim of finding patterns in the database provided by Mode Finance.

4.2.2 Neural Network implementation

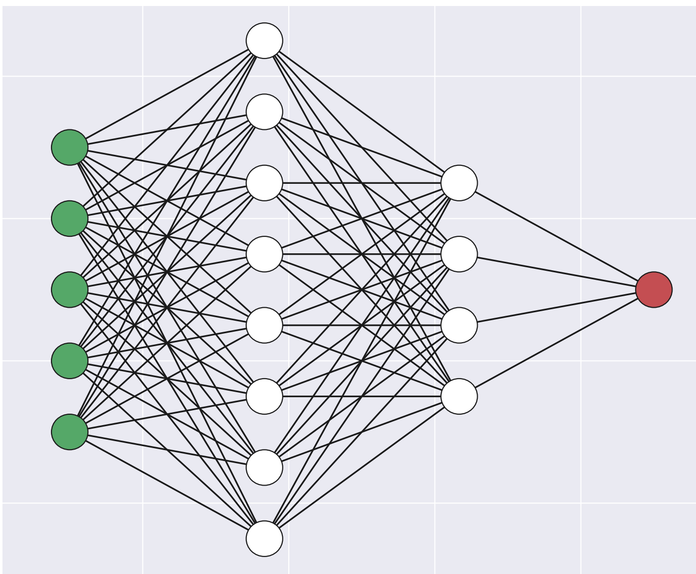

Figure 4.4: NN graphical representation, taken from [9].

Machine Learning (ML) is the ”study of computer algorithms that improve automatically through experience” [32]. From a mathematical point view, it is possible thanks to the Representer Theorems [39]. In an extreme simplification, these theorems state that any function, however complex, can be represented as a combination of simpler functions.

One of these structures, which we consider here, is a deep Neural Network, see Fig. 4.4. In this structure, inputs of the function are fed to a first layer of so-called ”neurons”, which compute linear combinations of them and then pass the output through an ”activation function”, feeding it to the next layer. At the final step a single output can be provided (regression or binary classification) or multiple ones (e.g. classification).

Here we look at the problem as a classification one: we have a large dataset composed of n items, with each item being described by a certain amount of features $x \in \mathbb{R}^m$ and a scalar label y. Machine Learning, in this context, means finding the function $f(\cdot)$ which ”explains” the relationship between x and y. This explanation occurs in probabilistic terms, so that when a test sample x is given, the prediction $\hat{y} = f(x)$ is the one with maximum likelihood of being correct, based on all the pairs of known samples (x_i, y_i) which include feature and label.

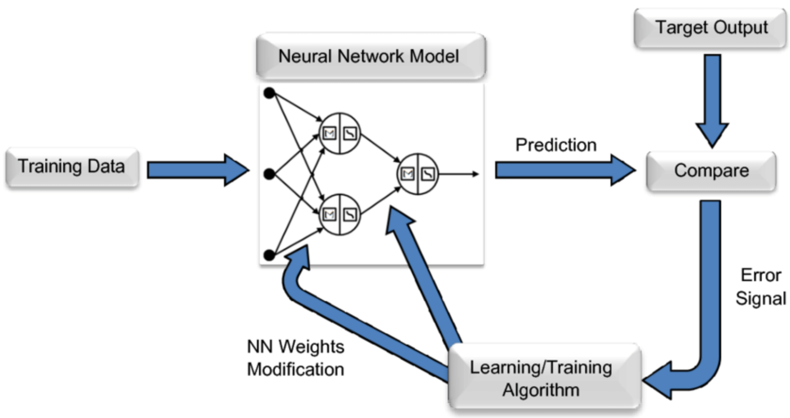

Figure 4.5: NN training

In order to reach the point in which the function $f(\cdot)$ is able to predict the label of new samples, a training process has to occur, graphically explained in Fig. 4.5. The function $f(\cdot)$ is a parametric function, depending on a set of p parameters $\theta = [\theta_1, \theta_2, ..., \theta_p]$ which are initially unknown, and have to be learned through the training process, so we write it as:

$$y = f_\theta(x) \tag{4.1}$$

Before training the NN, the available dataset is divided into a training set $(x_{tr,i}, y_{tr,i})$, for $i = 1, ..., n_{tr}$ and a test set $(x_{ts,i}, y_{ts,i})$, for $i = 1, ..., n_{ts}$, where $n_{ts} + n_{tr} = n$. The typical balance of training set and test set is such that between 15% and 30% of the overall dataset is used for the testing. The training takes several iterations, where ”batches” of data (x_b, y_b) are evaluated, and the parameters of the function (4.1) are updated via a gradient algorithm, in order to minimize a loss function, i.e. $\sum(y_b - \hat{y}_b)^2$, where $\hat{y}_b = f_\theta(x_b)$. The training is iterated multiple times with all the training set, with each iteration called "epoch". This is required because parameters θ have to be updated slowly, and the information has to be learned gradually. Theoretically, the learning phase could go on indefinitely, and the ability to explain the training set would keep on improving, however this would degenerate in over-fitting, where the capability to generalize over other instances from the same distribution would be compromised.

The characteristics of the function (4.1) depend on the underlying NN structure, which in our case is of the form illustrated in 4.4. It is a sequential Neural Network with linear layers and ReLU activation function (see [32] for details).

Application to Credit Rating Analysis

The problem of classifying the companies into credit rating classes, is transformed into a regression problem, assuming a linear progression from the lowest class D to the highest AAA, as a linear scale from 1 to 10 (ModeFinance adopts the standard used by S&P, Fitch in Table 3.1).

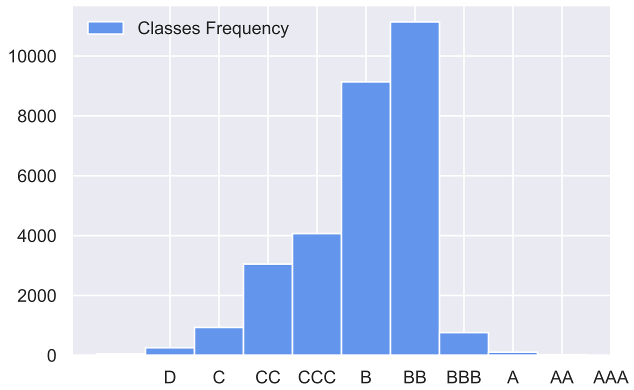

Figure 4.6: Frequency of credit rating classes in the available dataset.

As one can notice from Fig. 4.6, the distribution of samples across the different classes is considerably imbalanced. This is a significant problem when applying machine learning, where the information used for training should be as exhaustive and balanced as possible. For this reason, rating classes at the extremes (AAA/AA/A, and CC/C/D) are lumped together.

For the purpose of the analysis, a dynamic NN composed of a variable number of layers (maximum 5) has been created, based on the module PyTorch. Simulations showed that a structure with 2 hidden layers was sufficient for the available database, which has a limited complexity and dimensionality.

The most important aspects of pre-processing and training are the following:

- the Neural Network is trained and tested on the full database, which includes all items with at least one missing feature;

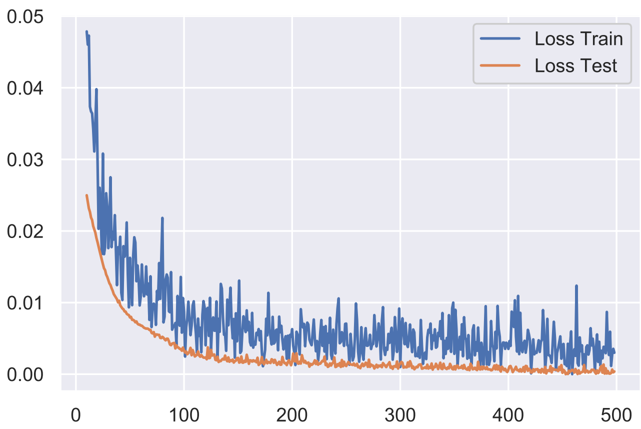

Figure 4.7: Training history: loss functions. Both loss functions are normalized with respect to their maximum value.

- missing features are filled with 0 (which is identical to having no information for a NN);
- features are Min-Max normalized along the columns to the range 0-1;
- in order to compensate for the imbalanced dataset, the Training Loss is computed using a weighted Mean Squared Error (MSE), where errors in classes with fewer samples (AAA/AA/A and CC/C/D) are weighted more:

$$L(x_b, y_b) = \frac{1}{n_{batch}} \sum_{i \leq n_{batch}} (w(y_i) \cdot (f_\theta(x_i) - y_i))^2$$

  where $w(\cdot)$ expresses the dependency on the correct credit rating class;
- contrary to the Training Loss, the Test Loss is a plain MSE, since the purpose is to properly classify all classes, without a priority on the accuracy distribution.

The evolution of the normalized training and test losses is shown in Fig. 4.7. The training is interrupted as soon as a clear stabilization in the prediction capability on the testing set is reached, to avoid over-fitting of the training set. This occurs approximately around epoch 450.

Prediction results

In order to assess the classificator's prediction capability, we define the following 2 kinds of accuracy, which will be used also in the evaluation of the other classification methods in the next sections:

- "Accuracy 0" ($Acc0$): Percentage of instances where the exact class is predicted;
- Accuracy 1 ($Acc1$): Percentage of instances where the exact class or neighboring class is predicted.

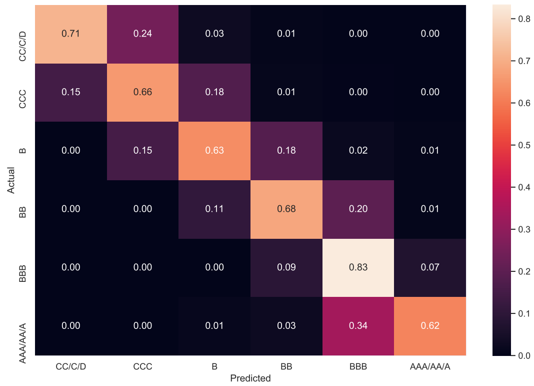

Figure 4.8: ANN: Normalized confusion matrix of the predictions on the test samples (sum along rows equal to 1). In this simulation, $Acc0 = 72.98\%$, $Acc1 = 98.8\%$, Average prediction deviation: 0.05

	Asset Side	Long Term Debt	Loans
Accuracy 0	71.05%	62.5%	79.2%
Accuracy 1	98.5%	100.0%	99.56%
Average prediction deviation	-0.03	-0.19	-0.02

The Neural Network regressor, which has been trained as described in the previous section, is evaluated against the test sample, which amounts to 20% of the total available dataset.

In order to have generalized results, the NN has been trained and tested multiple times with different splits Train/Test. $Acc0$ is in the 72%-74% range for all simulations, while $Acc1$ is always 99%, which is an excellent result.

The mean prediction error is always within the ± 0.05 range, which could be interpreted as if the NN has "learned" how to assign the correct class in average, without introducing bias (in statistical terms, $f_\theta(\cdot)$ is a correct estimator). The normalized confusion matrix of one simulation is shown in Fig. 4.8, while the numerical evaluation of the accuracy for the different types of missing features is included in Table 4.2.2. These values show that the predicting performance slightly deteriorates in case of missing data regarding the Long Term Debt, which leads the classifier to estimate a lower class (-0.19 average estimation). These results will be used as benchmark when evaluating the second method, which is not based on deep learning.

4.3 Random Forest Classifier

Random Forest (RF) classifiers are among the most used ones in Machine Learning, and have also been applied in several instances to the problem of estimating the credit rating of companies, see e.g. [41]. While a more exhaustive explanation of the principle behind a Random Forest classifier can be found in [12], here we will provide

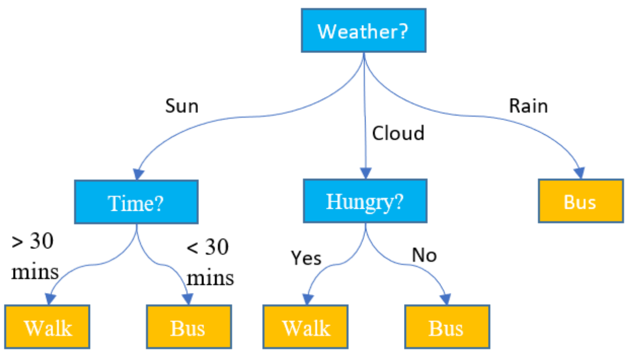

Figure 4.9: Decision Tree example: walking or taking the bus, depending on three variables (from [10]).

a high level explanation about how they work.

A Random Forest, as the name suggests, are based on Decision Trees, which are a basic tool to determine how features influence the belonging of a given item to a certain class. As shown in the example provided in Fig. 4.3, given different groups with respective labels, decision trees provide feature-based criteria to split the observed items, in order to have entries from the same group being as "similar" as possible among each others, and as "different" as possible from items classified with different labels.

In general, classifiers based on single decision trees tend to have poor generalization capabilities, since it is a tool which by nature over-fits the training data (prediction of a decision tree on the training sample is always 100% accurate). For this reason, the so-called "wisdom of crowds" is used, in order to exploit the simplicity of implementation of a Decision Tree, while at the same time overcoming the over-fitting problem.

A Random Forest comprises several single Decision Trees, having them operating as an "ensemble". The prediction of a Random Forest classifier is the result of a voting process in which all predictions of the individual Decision Trees are evaluated. The reason why RF is so powerful is that the underlying Decision Tree models are relatively uncorrelated among themselves, since they are trained on different batches of the same training dataset.

The principle behind this result is that different trees tend to negate individual errors, assuming there is not a general bias towards one direction, which could be a dataset problem, and would affect most classification methods. While an individual tree may be wrong, the majority of the others is likely to be right, so that in general the "ensamble" will produce the right response.

4.3.1 Application to the Available Database

In applying the RF classifier, it was decided to exploit the reduced database, in part based on the financial ratios presented in Section 3.2.1. The reasons behind this choice are mainly two:

1. a Random Forest classifier tends to perform particularly well when it is allowed to compare features against thresholds, since the underlying decision trees do not need to be particularly long in such instances. Using financial ratios therefore goes in this direction;

2. the data imputation step is based on the EM algorithm (see 4.4.1), which assumes the samples are all taken from a Gaussian Multivariate distribution. The level of complexity and computational costs increase dramatically with the dimension of the features, therefore using financial ratios helps also in this regard, since it allows to use less than half of the number of features.

For this reason, All financial ratios presented there are included in the database, when necessary using the approximation for the Net Income:

- Current Ratio
- Quick Ratio
- Cash Ratio
- Debt-to-Equity Ratio
- Return on Assets
- Return on Equity

In addition to these, two additional features are considered:

- Total assets, in order to account for the dimension of the company;
- Ratio between Tangible and Total Assets, to account for the presence of Intangible Fixed Assets.

Training Phase

Similarly to the ANN case, one of the most important issues in training the RF classifier is re-balancing the training set. In this case, though, it was found that grouping the credit rating classes at the extremes (AAA/AA/A, CC/C/D) was not sufficient. In fact, as previously shown in Fig. 4.6, the distribution of companies is much more concentrated in the central classes (B, BB), compared to others.

Moreover, in contrast to what was done for the ANN, only the fields with no features missing can be used for the RF classifier. This reduces considerably the pool of available entries, also in light of the fact that the ROA and ROE financial ratios required two consecutive years of Equity information to be computed, implying that the year 2016 could not have a single item used for the RF training.

The training set has been further re-balanced, in order to be able to have acceptable predictions also in case of rating classes with few samples. Classes were divided in ”frequent” (BB, BBB), ”common” (B, CCC) and ”unusual” (AAA/AA/A, CC/C/D), and the proportion in the samples belonging to each of those category was chosen via trial and error approach. In the presented results, the proportion used is 10 (frequent),

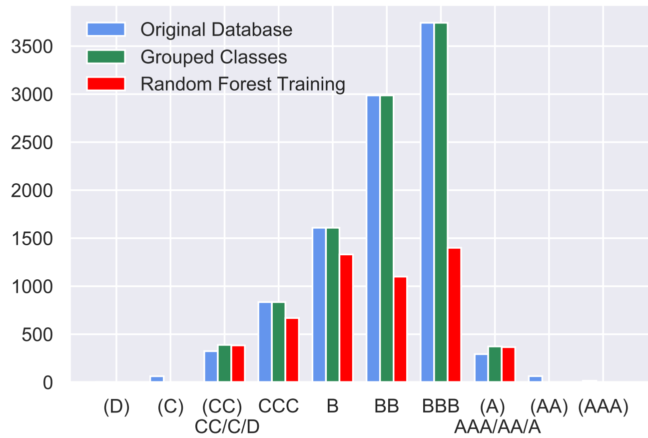

Figure 4.10: Distribution of samples across the Rating Classes, complete features only using the database with financial ratios.

8 (common), 3 (unusual), using 5250 randomly chosen samples from the complete portion of the database. In total, 53% of the available "complete" entries are used.

The distribution of 'complete' samples across classes, amounting to $\approx 1/3$ of the total, is represented in Fig. 4.10, where the frequency of entries in the different classes is shown for:

- Original Database (complete portion of the database)
- Grouped Classes (the one above where AAA/AA/A and CC/C/D are lumped)
- Random Forest Training-set.

Classification Results

The RF classifier is tested against all samples not used for the training. This test set is extremely unbalanced, due to the considerable amount of samples in the frequent classes not used for the training. Therefore these results are not very informative, and should not be used for a straight up comparison with the ANN predictions presented in 4.2.2.

That said, depending on the train-test split, the RF predicts correctly:

- 100% of the samples used for the training;
- $75\% - 80\%$ of complete samples not used for training ($Acc0$);
- $98\% - 99\%$ with one class of tolerance ($Acc1$).

These results are outperforming the NN when missing data is not considered, with a comparably low bias in the prediction (± 0.05). The normalized confusion matrix for one simulation is shown in Fig. 4.11.

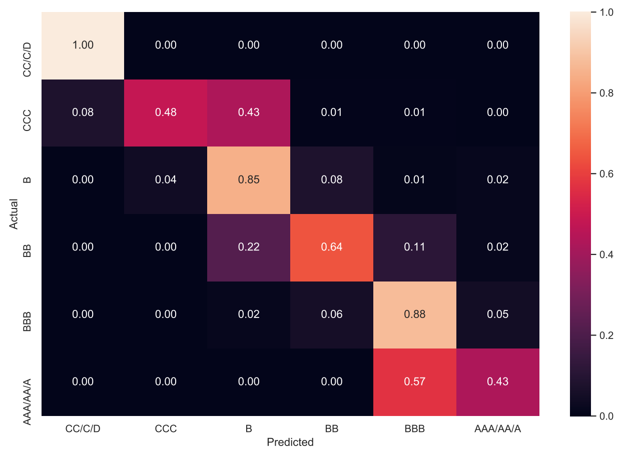

Figure 4.11: Normalized confusion matrix of the predictions on the test samples. In this simulation, $Acc0 = 76.72\%$, $Acc1 = 98.06\%$, Average prediction deviation: -0.03.

4.4 Features Imputation

The handling of large datasets with missing data is one of the main topics of interest in data science. Among several existing techniques, two different approaches can be identified to face this problem [36].

The first one is a "black box" approach, which exploits tools such as Deep Neural Networks, previously introduced in 4.2. This approach is very useful when there is strong correlation between the different entries and features, and is especially useful in cases when there is little knowledge on the exact relations existing between the data. The algorithm for credit rating estimation by Mode Finance exploits these tools as a part of their employed methodologies. The ANN approach, which we applied earlier, also exploits these properties, so that ANN could be applied directly to the dataset with missing features.

A second and radically different approach is to perform data imputation, by means of a statistical analysis of the dataset. Several methods exist in this regard: the most frequent (and most simple) ones include are those which use the mean, median, or most frequent value as imputation variable [38]. However, when the problems become more complicated, it might be more efficient to use more complex algorithm for this purpose. One of these, known as EM Imputation Algorithm [33], will be discussed in the next subsections, and used to perform data imputation on the available dataset of German companies, in order to find a suitable classification criterium.

4.4.1 Expectation Maximization (EM) Imputation

Hypothesis

In this work we use a procedure to impute missing components in each data point using the EM algorithm. This procedure assumes that for each entry (in our case financial ratios of a given company in a given year) the ratios are iid (independent and identically distributed random variables) extracted from the same normal multivariate distribution.

This assumption is made in spite of the fact that there is a clear correlation between the ratios of the same company over consecutive years. In addition, the information about the economic sector in which each company operates has to be neglected, since such information (one-hot encoded for training purposes) cannot be translated in a continuous numeric form.

i/j	**Feature 1**	**Feature 2**	**...**	**Feature j**	**...**	**Feature p**
Element 1	X_{11}	X_{12}				X_{1p}
Eelement 2	X_{21}	X_{22}				
...						
Element i	X_{i1}			X_{ij}		X_{ip}
...						
Eelement n	X_{n1}					X_{np}

Starting from these hypothesis, the available database X, excluding the credit rating information, is in the form reported in Table 4.4.1, where the hypothesis expressed above can be written as:

$$X_i = (X_{i1}, ..., X_{ip}) \sim N_p(\mu, \Sigma) \tag{4.2}$$

Moreover, we express the fact that certain features are missing by defining an $n \times p$ boolean matrix C, such that:

$$C_{ij} = \begin{cases} 1, & \text{if } X_{ij} \text{ is available} \\ 0, & \text{if } X_{ij} \text{ is unknown} \end{cases}$$

and starting from C, we finally define for each row i the complementary sets

$$A_i := \{j | C_{ij} = 1\}$$
$$U_i := \{j | C_{ij} = 0\}$$

Note that, in order to have sufficient information to perform the imputation of missing values on a line, at least one of the features have to be available, so that $A_i \neq \{\}$ for any i.

Starting from these conditions, the EM algorithm updates at each iteration the set of parameters $\theta = (\mu, \Sigma)$ which describe the multivariate distribution, until a certain convergence condition is met. To do so, a starting estimation of $\theta(0) = (\mu(0), \Sigma(0))$ also needs to be considered.

Algorithm Iteration

The core idea behind the EM algorithm is that the unknown (or unobserved) components are distributed (together with the available ones) according to the multivariate distribution (4.2). The purpose of the EM algorithm is to iteratively find the set of parameters θ which maximizes the likelihood of that distribution, given the observed data. The expected value of the parameters of the likelihood function, given the set of parameters θ identified at the previous step, can be written as:

$$Q(\theta|\theta(t)) := E_{\theta(t)}[\log L(\theta; X)|\boldsymbol{X_a} = \boldsymbol{x_a}] \tag{4.3}$$

$$\sum_{i=1}^{n} E_{\theta(t)}[logL(\theta; X_i)|X_{ia} = x_{ia}] \tag{4.4}$$

where the equality is allowed by the iid assumption, and we refer for simplicity to the set of observed feature as X_a.

The EM (Expectation Maximization) procedure is an optimization, in which at each step:

- E: the expectation function (4.4) of the (log-)likelihood of the dataset is computed;
- M: the new set of distribution parameters is obtained by maximizing (4.4): $\theta(t+1) = \arg\max_\theta(Q(\theta|\theta(t)))$.

The mathematical explanation of how this procedure works can be found in [33, 2]. The Python code used for the procedure has been derived from [3].

4.4.2 Classification Results on the Imputed Dataset

The procedure to test the imputed database on the trained RF classifier comprises two preliminary steps:

1. the entire database is used for the imputation phase, according to the procedure described in 4.4.1;
2. the entries which were originally used for the training of the RF classifier are dropped (in this way the test phase doesn't include any sample already used for the training).

Once the training entries have been pruned, the test set is ready to be evaluated. Note that, given the fact that only $\approx 1/6$ of the dataset was used for the RF classifier training, the test data will be considerably larger than in the ANN case, and therefore less variance in the results is expected when changing train-test splits.

The normalized confusion matrix for one simulation is show in fig. 4.12, with numerical analysis of the accuracy for different types of missing features in Table 4.4.2.

In general, Accuracy scores are not much worse than in the NN case, but the average prediction error is considerably higher (0.16). That could mean that the NN "learned" how to assign the correct class in average, without introducing bias.

Analyzing Table 4.4.2, one can see that, compared to the ANN case, also with the second method the predictions generated the largest percentage of mis-classifications when Long Term Debt information is missing. In this case, though, there is a considerably larger proportion of companies which were classified in a higher class

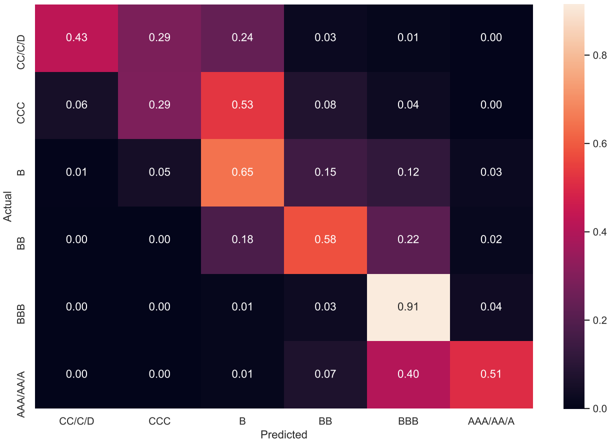

Figure 4.12: Normalized confusion matrix of the predictions on the test samples (sum along rows equal to 1). In this simulation, $Acc0 = 68.64\%$, $Acc1 = 94.59\%$, Average prediction deviation: 0.16

	Asset Side	Long Term Debt	Loans
Accuracy 0	66.82%	56.25%	66.62%
Accuracy 1	93.05%	90.62%	92.56%
Average prediction deviation	0.09	-0.22	0.26

(keeping in mind that the performance of the NN regressor is expected to perform better, due to the fact that a portion of the evaluated inputs were also used for the training). In general, one can notice how, although the accuracy scores are not too far away from the NN case, the classifications tend to have a considerable bias when the imputed data is analyzed.

4.5 Considerations

Comparing the performance of the two methods, one can immediately notice how the NN performs better than the combination of EM imputation and RF classifier. By construction, a properly trained NN is able to capture all the information and correlations within the database, although it is not trivial to explain them due to its black-box nature.

The EM-RF method is considerably more complicated, consisting of 4 main steps:

1. definition of financial ratios and approximations required to reduce the database dimensionality;

2. training of the RF classifier (simpler than the NN, but very efficient on low-dimensional datasets) on the complete portion of the database, to capture relations between financial ratios and credit rating;

3. imputation of missing values via the EM algorithm;

4. prediction via the RF of the imputed dataset, after removing the training data.

While the prediction output of the NN is excellent ($Acc0 =\approx 75\%, Acc1 =\approx 98\%$), the EM-RF method comes close in both accuracy measures ($Acc0 =\approx 69\%, Acc1 =\approx 95\%$). Moreover, when considering the prediction capability of the two methods in the cases when information was missing, we found out that the classification is in line with the average, in the cases where Asset related or Loans information are missing, while performance worsens considerably when Long Term Debt information is missing.

From these results, some general considerations can be made. The ANN is a powerful tool, which in these testing conditions has also the advantage of being able to exploit a much larger training set: the Neural Network uses 80% of the entire database for training, while the Random Forest is trained only on 17% of the full database. This is due to two reasons: (i) it is not possible to train the RF on non-verified imputed data, (ii) the training-set has to be balanced at least partially. Given the different proportion of information used, EM-RF displays a good performance, and seems to validate the strong hypothesis made before the analysis:

- the choice of financial indicators chosen carries much of the information included in the entire database;
- the NACE classification helps in the classification of the Credit Classes, although it is not as crucial in explaining the multivariate distribution of features based on financial ratios;
- the hypothesis itself of entries as iid from a Gaussian multivariate distribution seems plausible, although we know that at least entries from different years, related to the same company, are significantly correlated.

Beyond these conclusions, this work suggests a number of possible directions for the investigation on the best way to deal with missing data for the credit rating of SME companies:

- evaluating alternatives to the replacement of NaNs with 0 in the NN training: at the moment the different kinds of missing data (no data vs. 0) are considered in the same way, also thanks to the ANN properties. However, in certain situation 0 should carry a lot of information, e.g. the fact that a company has no long term debt. A deeper analysis of this aspect represent the most interesting direction in which to move the research;
- in this work companies' balance sheets from different years are considered independent entries, which neglects the fact that the history of that specific company is included in the database. Therefore a different kind of training for the NN should be considered, in which the credit rating evolution over multiple years is factored in, by means of 2D input layers, or, if large time window is available (at least 10 years), Convolutional and/or Recurrent NN;
- exploring imputation methods different from EM, e.g. auto-encoders (which are also based on ANNs);

- investigating why the method based on EM-RF introduces bias in the rating estimation, whereas the ANN does not: this at the moment represents the most glaring weakness of the AM-RF method.

Chapter 5

Conclusions

Actors on the economic and financial stage have realized the importance of having widespread access to information regarding the creditworthiness of companies at all levels. Examples of this are the recommendations on banking laws and regulations stated in Basel II [23]. Probability of Default (PD) models beneficiate considerably from an increase in transparency and an improved harmonization of the accounting standards.

As it emerged from the sources analyzed in Chapter 2, in Germany the regulation regarding the publication of financial records is not yet totally harmonized with the EU indications. German authorities still have an ambiguous stance towards SMEs, which are allowed to do their bookkeeping according to the European IRFS standards, but at the same time are required to comply to the German GAAP when dealing with public administration offices. Small companies benefit from the fact that they are only required to make their Balance Sheet public, while they do not need to present the Income Statement. Moreover, there is considerable freedom on their behalf to not disclose sensitive information in their Balance Sheets, e.g. the distinction between short term loans and long term debt.

The duty to publish more complete books, would be beneficial mainly because a larger number of ”drivers” (i.e. financial ratios) would be available in order to determine PDs. As we described in Chapter 3, only few financial indicators are available at the moment for the large population of German SMEs, with a lack of key fundamentals explaining Efficiency, Profitability and Market Value, which are important aspects to consider when evaluating a firms creditworthiness. In this context, the duty of publication of a basic Income Statement would make it possible to vastly improve regression models for the estimation of credit scores.

In the final Chapter 4 we have tried to provide an analysis of how credit rating can be performed, in spite of the lack of complete data, by means of modern Machine Learning techniques. We have used a database provided by ModeFinance, assuming it has the ”correct” information regarding companies’ credit scores. With these premises, a sequential linear Artificial Neural Network (which is a ”black-box” tool) shows great capabilities of identifying the patterns in the dynamics between balance sheet accounts and credit rating. In particular, it displays the property of being a ”correct” predictor, which in statistical terms indicates an estimator with no bias. ANNs of this type are well suited to handle missing data, although at the same time they have the drawback of considering a field equal to zero the same as an empty one. This aspect is problematic since companies not disclosing their debt are treated in the same way as companies with no debt.

A second approach, more statistical, shows that it is reasonable to assume that an underlying Gaussian multivariate distribution describes the dataset, in its reduced form where financial ratios are considered. However, the technique used to perform the so-called "data imputation" is computationally intensive and doesn't scale up well in case of datasets of growing dimensionality. This second approach displays a good performance, although key missing information in the Long Term Debt causes the classifier to have a positive bias in the computation of the credit scores. This is probably due to the fact that ModeFinance's estimations are more "conservative", when key data are missing.

Bibliography

[1] Accountingtools, segment reporting. https://www.accountingtools.com/articles/what-is-segment-reporting.html. Accessed: 2020-07-01.

[2] Em imputation algorithm, junkyu park github page. https://joon3216.github.io/research_materials/2019/em_imputation.html. Accessed: 2020-06-29.

[3] Em imputation: Python implementation, junkyu park github page. https://joon3216.github.io/research_materials/2019/em_imputation.html. Accessed: 2020-06-29.

[4] Esma, official website. https://www.esma.europa.eu/. Accessed: 2020-08-14.

[5] Eurostat, structural business statistics, sme. https://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/sme. Accessed: 2020-06-30.

[6] Ifrs foundation. https://www.ifrs.org/. Accessed: 2020-06-30.

[7] Investopedia, ifrs. https://www.investopedia.com/terms/i/ifrs.asp. Accessed: 2020-07-01.

[8] Investopedia, sme default probabilities. https://www.investopedia.com/articles/financial-theory/11/default-probabilities-small-medium-corporations.asp. Accessed: 2020-08-14.

[9] Nn, picture source. https://www.researchgate.net/profile/Syariful_Shamsudin/publication/299390844/figure/fig5/AS:650029635235894@1531990550259/The-representation-of-neural-network-training-process.png. Accessed: 2020-08-14.

[10] Sql shack,decision trees. https://www.sqlshack.com/microsoft-decision-trees-in-sql-server/. Accessed: 2020-08-13.

[11] Topvendors, modefinance. https://www.topvendors.it/en/fintech/modefinance. Accessed: 2020-08-14.

[12] Towardsdatascience, understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed: 2020-08-13.

[13] Wikipedia, bondcredit rating. `https://en.wikipedia.org/wiki/Bond_credit_rating`. Accessed: 2020-08-11.

[14] Wikipedia, one-hot encoding. `https://en.wikipedia.org/wiki/One-hot`. Accessed: 2020-08-13.

[15] Bao, W., Yue, J., and Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944.

[16] Bartels, B. and Weder di Mauro, B. (2013). A rating agency for europe a good idea?

[17] Bassemir, M. (2018). Why do private firms adopt ifrs? *Accounting and Business Research*, 48(3):237–263.

[18] Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111.

[19] Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

[20] Cavalcante, R. C., Brasileiro, R. C., Souza, V. L., Nobrega, J. P., and Oliveira, A. L. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55:194–211.

[21] Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, 83:187–205.

[22] Ciampi, F. and Gordini, N. (2008). Using economic-financial ratios for small enterprise default prediction modeling: An empirical analysis. In *2008 Oxford Business & Economics Conference Proceedings, Association for Business and Economics Research (ABER)*, pages 1–21.

[23] Decamps, J.-P., Rochet, J.-C., and Roger, B. (2004). The three pillars of basel ii: optimizing the mix. *Journal of Financial Intermediation*, 13(2):132–155.

[24] Eurostat, N. (2008). Rev. 2–statistical classification of economic activities in the european community. *Office for Official Publications of the European Communities, Luxemburg.*

[25] Haller, A., Ernstberger, J., and Froschhammer, M. (2009). Implications of the mandatory transition from national gaap to ifrs—empirical evidence from germany. *Advances in Accounting*, 25(2):226–236.

[26] Harper, S. (2011). Credit rating agencies deserve credit for the 2007-2008 financial crisis: An analysis of cra liability following the enactment of the dodd-frank act. *Wash. & Lee L. Rev.*, 68:1925.

[27] Horrigan, J. O. (1966). The determination of long-term credit standing with financial ratios. *Journal of Accounting Research*, pages 44–62.

[28] Lu, H.-M., Tsai, F.-T., Chen, H., Hung, M.-W., and Li, S.-H. (2012). Credit rating change modeling using news and financial ratios. *ACM Transactions on Management Information Systems (TMIS)*, 3(3):1–30.

[29] Mattarocci, G. (2013). *The independence of credit rating agencies: How business models and regulators interact*. Academic Press.

[30] Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

[31] Michalkova, L. and Kramarova, K. (2017). Capm model, beta and relationship with credit rating. In *Advances in Applied Economic Research*, pages 645–652. Springer.

[32] Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

[33] Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

[34] Moya, S., Perramon Costa, J., and Constans, A. (2005). Ifrs adoption in europe: the case of germany.

[35] Nehrebecka, N. and Polski, N. B. Probability-of-default curve calibration and validation of internal rating systems.

[36] Nelwamondo, F. V., Mohamed, S., and Marwala, T. (2007). Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*, pages 1514–1521.

[37] PWC (2018). *Doing Business and Investing in Germany*. PricewaterhouseCooper.

[38] VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.".

[39] Wahba, G. and Wang, Y. (2014). Representer theorem. *Wiley StatsRef: Statistics Reference Online*, pages 1–11.

[40] White, L. J. (2010). Markets: The credit rating agencies. *Journal of Economic Perspectives*, 24(2):211–26.

[41] Yeh, C.-C., Lin, F., and Hsu, C.-Y. (2012). A hybrid kmv model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, 33:166–172.

[42] Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480.