Bachelor's Degree in Business Analytics

Bachelor's Final Project

Development of a default prediction model for Spanish SMEs based on publicly available information

Author
Nicolás Oriol Guerra

Supervised by
Carlos Bellón Núñez-Mera
Antonio Uguina Zamorano

Madrid
2021 -2022

I declare, under my responsibility, that the presented Project titled

Development of a default prediction model for Spanish SMEs based on publicly available information

in the Engineering School - ICAI, Universidad Pontificia Comillas in the 2021/22 academic year is my own work, original and unpublished, and it has not been previously submitted for other purposes.

This project is not plagiarism of any other work, neither totally nor partially and all contributions by other authors have been dully referenced.

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Development of a default prediction model for Spanish SMEs based on publicly available information

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2021/22 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Nicolás Oriol Guerra Date: May 26th 2022

Project's delivery authorized
Autorizada la entrega del proyecto

THE PROJECT'S SUPERVISOR
EL DIRECTOR DEL PROYECTO

Carlos Bellón Núñez-Mera Date: June 9th 2022

Bachelor's Degree in Business Analytics

Bachelor's Final Project

Development of a default prediction model for Spanish SMEs based on publicly available information

Author
Nicolás Oriol Guerra

Supervised by
Carlos Bellón Núñez-Mera
Antonio Uguina Zamorano

Madrid
2021 -2022

Contents

List of Figures

List of Tables

1 Introduction

1.1 SME environment in Spain

SME stands for small and medium-sized enterprises (PYME in Spanish). Businesses may be classified as SMEs according to the definition provided by the European Union (*SME Definition*, 2022). This classification is based on three factors: staff headcount, turnover, and balance sheet total. According to these variables, businesses are labeled as *Micro*, *Small* or *Medium-Sized*. If they do not meet the requirements, the business is considered to be *Big* (non-SME). This classification method is further detailed in section 4.3

At a European level, SMEs represent 99% of all businesses. In Spain, numbers are similar with an even greater share of enterprises being considered SMEs - up to 99.9% (*Retrato de la PYME*, 2022). Overall, in January 2022 (the latest available data) Spain had just under 2.93 million businesses (down from 3.37 in January 2021). Out of these, 2.92 million (3.36 in January 2021) were considered SMEs (*Retrato de la PYME*, 2022), (*Cifras PYME Enero 2022*, 2022). Most SMEs are associated with the service industry - examples include restaurants, hotels, and bars - and are mostly concentrated in four Comunidades Autónomas (CCAA) [1]. Cataluña, Madrid, Andalucia and Comunidad Valenciana contain 61% of all Spanish businesses. Figure 1 shows how these 2.92 million SMEs are distributed based on number of employees.

Overall, SMEs represent a crucial piece of any country's economy. This is especially true in the Spanish case where 65% of the country's GDP and almost 60% of the country's employment are generated by these small businesses (*La PYME Española y el reto del crecimiento*, 2022), (*Cifras PYME Enero 2022*, 2022).

In Spain, SME's contribution to employment figures is 5.3 percentage points higher than the EU average. Moreover, the micro-sized SMES constitute the highest contribution to national employment: 38.7% against an EU average of 29.8%. (*Marco Estratégico en Política de PYME 2030*, 2021). Tables 1 and 2

[1] Highest level of political and administrative-territorial division in Spain. The Spanish territory is divided in 17 CCAA and 2 Ciudades Autónomas (Ceuta and Melilla)

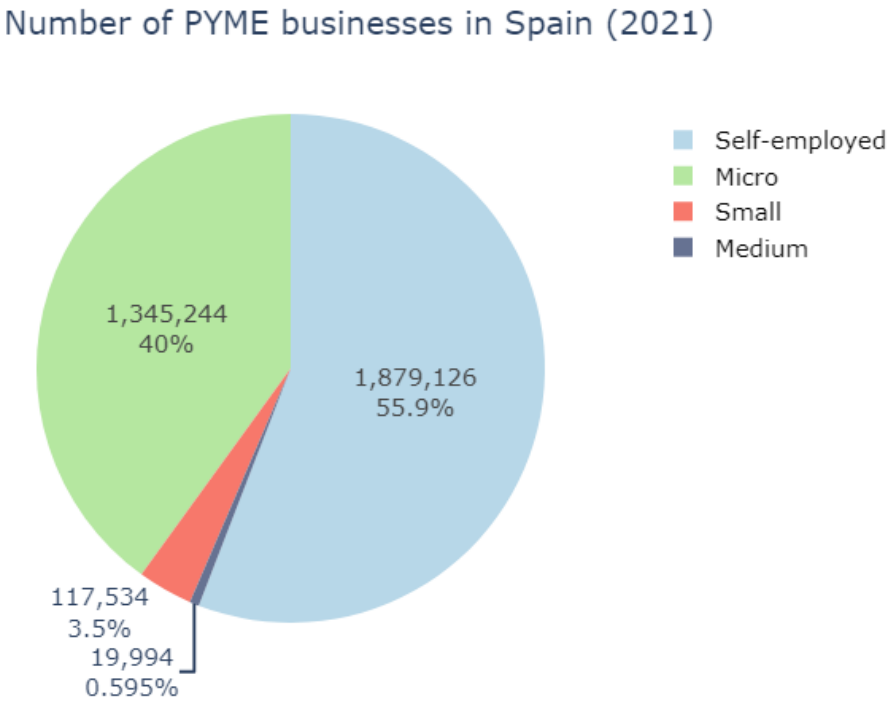

Figure 1: Number of SME businesses by size (*Retrato de la PYME*, 2022)

show the distribution of the number of companies by size and sector within the Spanish economy (*Retrato de la PYME*, 2022). The crucial takeaway from these tables is that SMES represent an essential piece within the Spanish market.

Sector\Size	Self-employed	Micro	Small	Medium	Large
Industry	3.6%	7.0%	20.9%	22.1%	21.8%
Construction	13.1	11.4	14.2	7.0%	2.8
Commerce	18.5	25.6	16.1	15.0	14.1
Rest of services	64.8	56.0	48.8	55.9	61.3
Total	100%	100%	100%	100%	100%

Table 1: Percentage of businesses in each economic sector by company size

For all these reasons, it is understandable that special attention and resources from both public and private parties are SME-oriented. On one hand, the Spanish Government and European Union have special funds and aid programs tailored for

Size\Sector	Industry	Construction	Commerce	Rest of services
Self-employed	35.4%	58.9%	48.6%	59.7%
Micro	49.0	36.8	48.2	36.9
Small	12.8	4.0	2.7	2.8
Medium	2.3	0.3	0.4	0.5
Large	0.5	0.0	0.1	0.1
Total	100%	100%	100%	100%

Table 2: Percentage of companies of each size by sector

SMEs (*Guía del Usuario sobre la Definición del Concepto de PYME*, 2016). On the other hand, banks and other financial institutions must have special services and models designed especially for them.

This project is focused on this second area. The goal of this study is to provide a predictive model that computes the probability of default of a SME given its annual accounts. This model will be built from scratch using only public information provided by INFORMA "INFORMA Filial de Cesce líder en el suministro de Información Comercial, Financiera, Sectorial y de Marketing de empresas y empresarios". This model and its inputs will be further detailed in section 1.3.

1.2 Historical trends. The effects of the financial crisis and Covid pandemic

The present study takes into account data from the 2008-2020 period. This is a rather particular period in the economic landscape. It comprises the aftermath of the 2007-08 Global Financial Crisis and the first year of the COVID-19 pandemic.

We will briefly lay out the effects of the financial crisis on SMEs. By doing this, we can provide a brief compilation of its most significant consequences on SMEs. The COVID-19 pandemic will not be contemplated in this study although its effects will most likely be felt in the coming years.

SMEs were one of the sectors that were most negatively affected by the 2008 crisis. The crisis hit SMEs by depriving them of financing opportunities. Although a sharp decline in demand also contributed to their troubles, lack of access to financing played a key role in the sector's problems. As stated in Hernández

and Buil Vilalata, 2012, newly-constituted SMEs are mostly self-financed. As their operating volume increases, their growth is usually supported by external financing. This secondary source of financing usually comes in the form of debt (rather than contributions to equity) and the most common lenders are banks. As financing became more costly in the years following the 2007-08 housing market crash, SMEs were strapped for liquidity. (*El número de pymes que se declararon en quiebra en España se ha triplicado desde el comienzo de la crisis*, 2022) states that the number of loan petitions that were turned down rose by five percentage points during the crisis. This difficulty in credit-accessibility was mainly due to two reasons: information asymmetry and lack of scale (Hernández and Buil Vilalata, 2012).

SMEs are at a natural disadvantage when negotiating with financial institutions. The latter may have access to information regarding the business's activities and usually have the upper hand during the negotiations. This asymmetry is inevitable. However, it may be considered beneficial to the system as a whole because it theoretically enables a more efficient allocation of capital by financial institutions.

Lack of scale is also an intrinsic and insurmountable difficulty that SMEs face when accessing capital markets. Their credit petitions are smaller and riskier than those of bigger players as they have less collateral and worse guarantees overall.

The combination of all of these factors, and the widespread effects of the financial crisis had devastating consequences for the SME environment. Notably, shareholder returns suffered a sharp decline from their 2004 peak of 17.48% down to their lowest point in 2013 of 1.96%. This effect was more spread out over time and less severe for bigger corporate entities (Blanco Ramos, Fernández Blanco, and Ferrando Bolado, 2016).

The economic crisis motivated by the COVID pandemic is proving to have significant consequences across all sectors. The SME sector has been one of the hardest hit because of its vulnerable position motivated by the small-sized businesses, low capabilities for digital transformation and high weight of service-related products. In a similar manner to the 2008 crisis, SMEs have taken the brunt of

the blow and their metrics are worse than those of larger corporations (Blanco et al., 2021). However, COVID is not taken into account in the current study. This exclusion was done for two main reasons: the crisis is still ongoing and its effects are still uncertain, and the economic data for 2021 - a necessary starting point for any study - is not fully available at the time of writing.

Overall the period under consideration (2008-20) is very interesting from an economic point of view as it encompasses the full aftermath of the 2008 financial crisis, the following recovery and bonanza of the late 2010s and the first year of the COVID pandemic. A general idea of this economic cycle can be inferred from the GDP growth shown in figure 2

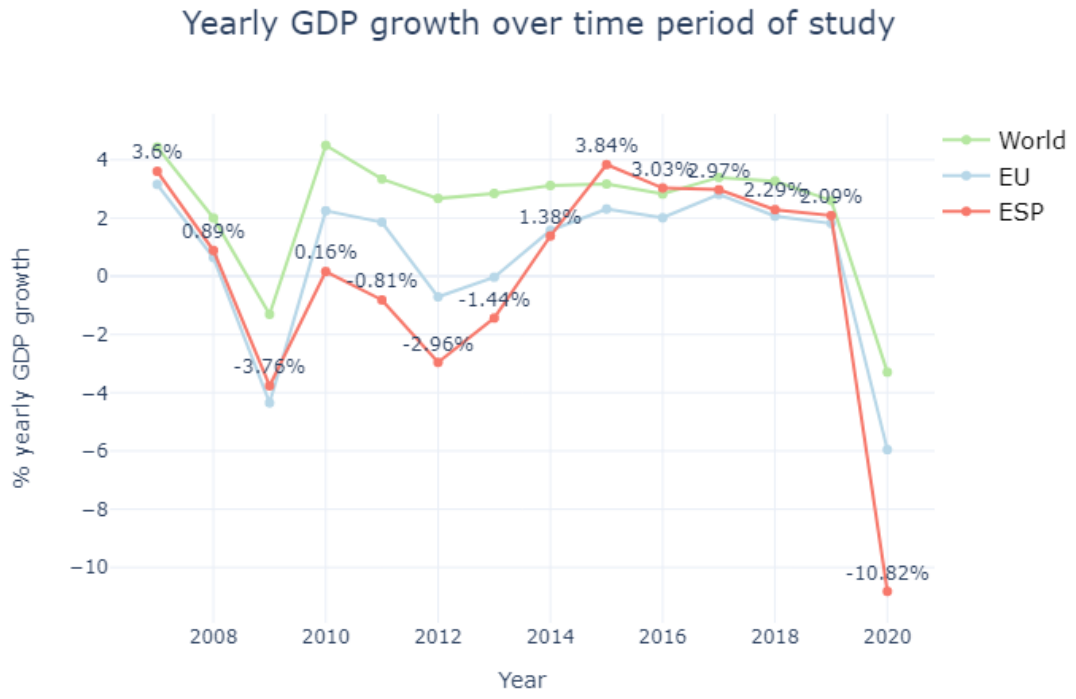

Figure 2: Yearly GDP growth (*GDP growth, annual %*, 2022)

1.3 Focus of the study

This study aims to provide a model that calculates the probability of default of a given SME for a time horizon of 1 to 4 years. This model will be trained with public data from INFORMA, across all years of the study. After this introduction, The paper is structured as follows. In section 2, we discuss previous work in this field. The baseline model used for comparison purposes will be Altman's model (Z-score). Details detailing the new angle of approach will also be discussed in this section. Section 3 describes the input data that was used to train the model. We provide some initial considerations and description of the data tables provided by INFORMA to then perform an Exploratory Data Analysis and arrive at some initial relevant conclusions. Section 4 describes the transformations and filters that were applied to the raw data provided by INFORMA. All decisions regarding data exclusions and treatment are justified and explained in detail. We also provide a detailed list of all calculated ratios. The final section (5 lays out the obtained model and explains the steps that have been taken to build it. It also presents several conclusions that can be extracted from the predictions. This conclusions are also represented in 6

2 State of the art

Edward I. Altman was the pioneer that paved the way for the academic study of bankruptcy prediction using modelization techniques. In 1968, he published a paper titled *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy* which aimed to defend the approach of "ratio analysis" as a valid way to assess probabilities of corporate bankruptcy (Altman, 1968).

2.1 Predicting default

In 1968, Altman established a discriminant model that allowed for the classification of businesses between bankrupt and non-bankrupt firms. He did this to improve upon the traditional techniques of ratio analysis that were common at the time. Financial ratios can provide useful insights regarding a company's financial structure and overall position. Ratios can measure profitability, solvency or liquidity among other factors, and they allow analysts to compare companies of different sizes and even across different industries.

Altman spotted a problem regarding the possibility of different ratios indicating opposing signals. Take, for example, a company with a low current ratio and high return on assets (ROA). If the current ratio (current assets/current liabilities) of a firm is too low, it indicates that the company may be experiencing some problems regarding the payment of short-term obligations. The ROA (net income/total assets) ratio shows how profitable a company is in comparison to its total assets.

This example company is sending mixed signals by indicating difficulties in meeting short-term creditor demands whilst hinting at high profitability. These are opposing signals and - in more intricate cases - may be difficult for investors to analyze and interpret. Altman proposes a multiple discriminant analysis (MDA) to tackle this problem and propose an alternative solution.

MDA provides a useful and easy-to-interpret method to separate and classify firms based on their most relevant characteristics. Altman's result consists of a single equation that yields a unique numerical result for every input. The equation (or discriminant function) is shown below (Altman, 1968):

$$Z = 0.012 * X_1 + 0.014 * X_2 + 0.033 * X_3 + 0.006 * X_4 + 0.999 * X_5$$

where:

X_1 = Working capital / Total assets
X_2 = Retained earnings / Total assets
X_3 = Earnings before interest and taxes / Total assets
X_4 = Market value equity / Book value of total debt
X_5 = Sales / Total assets
Z = Overall Index

Each of the proposed ratios that Altman uses in his model provides information about a different aspect of corporate activities (Guest, 2021). X_1 compares working capital (current assets - current liabilities) with total assets and it is a measure of company liquidity. X_2 measures retained earnings against total assets. This ratio is an indicator of age - younger companies tend to have lower retained earnings (Altman, 1968) - accumulated profitability and financing sources. X_3 is a profitability metric that calculates the return on assets (ROA) based on EBIT. X_4 is a strong indicator of long term solvency. The final ratio X_5 measures the firm's asset turnover.

2.2 Historic models

Altman's MDA model developed in 1968 has several shortcomings. The three main ones are sample selection, historic validity, and data availability. Regarding sample selection, Altman selected 66 corporations with 50% of them being bankrupt. The mean asset size of the selected companies is $6.5 million (the smallest having $0.7 million and the largest $25.9 million). This does not apply to all cases, as the distribution of bankrupt/non-bankrupt firms is significantly lower than 50% (in the case of the studied Spanish SMES is 3%). Moreover, the mean asset size is

also lower (5.62 million euros [2]). In the case of historic validity, Altman's model was developed over 50 years ago. It is expected that the weights and considerations must have changed over time. This is also stated by Altman himself in Altman et al., 2017. The third problem is data availability. The ratio X_4 in Altman's model uses the market value of equity as the numerator. This data is only available for public companies and thus, it does not apply to the present case. This issue is also addressed by the author in Altman et al., 2017 and Malakauskas and Lakstutiene, 2021.

All in all, the authors in Altman et al., 2017 update Altman's model based on six starting hypotheses. They also particularize the results for individual European countries. For example, in the case of the Spanish environment, they manage to significantly improve the benchmark by adding size-related variables to the prediction model.

2.3 The Spanish case

The final step in our study of previous works involves a more detailed look at more contemporary methods. More applicable examples for Spanish SMEs can be found in Malakauskas and Lakstutiene, 2021, and Camacho-Miñano, Segovia-Vargas, and Pascual-Ezama, 2013.

In Malakauskas and Lakstutiene, 2021, the authors use artificial neural networks as a predictor for SME financial distress. This approach will also be attempted with the current data. Another consideration made by the authors is the possibility of including additional variables related to multiple-year performance, not just using the one-year snapshot of the company's accounts. This paper is useful for the Spanish case because it is applied to small and medium enterprises using a wide array of analysis techniques.

The second paper that is more closely related to the current study is Camacho-Miñano, Segovia-Vargas, and Pascual-Ezama, 2013. In this case, the authors are

[2]This average is taken across the 2008-2020 period. No currency translation has been performed as several factors including inflation have to be taken into account and the final result is not relevant. It is enough to state that the used sample is sufficiently different (even the currency!) to deserve further consideration.

specifically looking at Spanish firms, which is especially relevant for our case. Overall, they determine that sector, size, ROA, stakeholder structure and liquidity can be strong predictors of a firm's survival capacity.

2.4 Improvements and build-up over previous works

Taking into account the previous work mentioned above, we will now define the desired aim of the current study and explain how it contributed to the available literature. Our goal is to develop a model that predicts SME bankruptcy procedures before they happen. In our case, we consider any court order related to insolvency proceedings to be the detonating cause of this bankruptcy procedure. Furthermore, we will also contemplate a secondary measure of financial distress based on Keasney, Pindado, and Rodrigues, 2015. In short, a firm is considered to be in financial distress if it meets the three following conditions:

1. EBITDA is less than financial expenditure for two consecutive years
2. Net worth/financial debt is lower than 1 for the current year
3. Net worth has negative growth between the previous and the current year

To develop this prediction model, we have the data described in section 3 which has been treated according to the procedures shown in section 4 to arrive at the final model shown in section 5.

Initially, we will define Altman's model as the baseline to outperform. This is a rather unfair comparison as Altman's training data only included public companies and the overall macroeconomic environment was very different to the one presented in our data. Nevertheless, when applying Altman's model to current companies results are not promising. SMEs are not publicly traded and thus, some of Altman's ratios don't even apply. When testing Altman's model on 2019's SME data, the results were extremely poor: 73.6% of the companies were classified to be in financial distress. Accuracy was below 45% (even for a highly imbalanced

dataset like this one) and precision was 16%. The results were similar when applying more modern versions of Altman's model. These results justify the need for the training of a new classifier for the particular case at hand.

The proposed roadmap for achieving some improvement over Altman's model is performed in two steps. First, we will redefine a new discriminant analysis model using the current data as our training set. Secondly, several different machine learning models will be trained on the available data and compared against each other to obtain the best-possible performing model. Models will be compared with each other based on their predictive results. The main metric, in this case, will be recall. Recall indicates the percentage of positive (bankrupt) cases that are correctly identified. For the case at hand, it is crucial to identify as many bankruptcies as possible even at the possible expense of some false positives. Accuracy (percentage of correct predictions) is not a good metric for this problem due to the high class imbalance - most companies do not go bankrupt. However, a possible alternative could be precision (accuracy of positive predictions) which prioritizes not having any false positives.

3 Description of available data

3.1 Data overview and initial considerations

The data available to us was presented by INFORMA in four different tables: `GENERAL_DATA`, `BALANCE_SHEET_HEADERS`, `BALANCE_SHEET_DETAIL`, and `BANKRUPTCY_DATA`. Figure 3 shows how the data is structured within these tables.

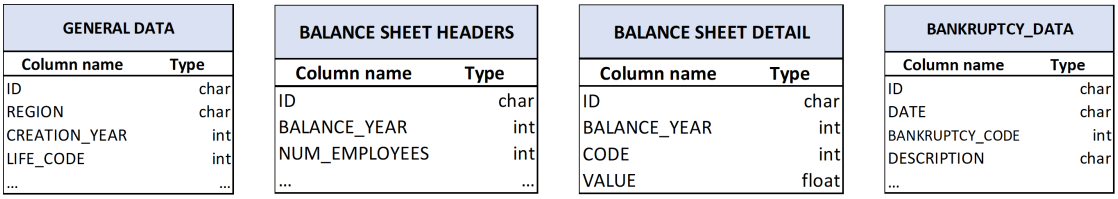

Figure 3: Tables provided by INFORMA

There is a total of 1.2M (million) SMEs in the dataset. The time period comprised within this dataset is from 2008 to 2020. Not all SMEs have data for every year. However, for the years where data is available for a company, there can be up to 88 different balance sheets and P&L values (CODE, VALUE) for it.

Overall, we have 1.2 million companies distributed over 14 reporting years in our dataset. Each company has presented an average of 8.11 balance sheets over the studied time period. Some companies present more than one balance sheet in a given year. The overall bankruptcy rate is 3.62% which amounts to a highly imbalanced datasheet. In section 4 we show each of the steps taken to process, organize, and order the data. In the section 5 we also describe how the imbalance in classes is considered when building the model.

3.2 Exploratory data analysis

In this section, we will comment on several findings and conclusions that can be extracted by analyzing the four tables presented above.

3.2.1 Table 1: General Data

This table has one row for each of the 1.2M companies. It provides a general overview of the individual businesses regarding the creation year, its current situation (`LIFE_CODE`), also information regarding its business sector. Figures 4, 5, and 6 illustrate some of the following conclusions.

- The region with the most registered companies is Madrid (223k) followed by Barcelona and Valencia (165k and 67k respectively). On the other hand, Soria is the region with the least registered companies with just over 2k. The cities of Ceuta and Melilla have even fewer companies.
- Of all the companies, 428k have a `LIFE_CODE` that is different from "ALIVE" (code 0). This does not imply bankruptcy as some companies may become inactive over time or be sold and may appear as "INACTIVE".
- Most company sizes are "Micro" (according to the number of employees only), and there are no NULL values or repeated ID in this table.

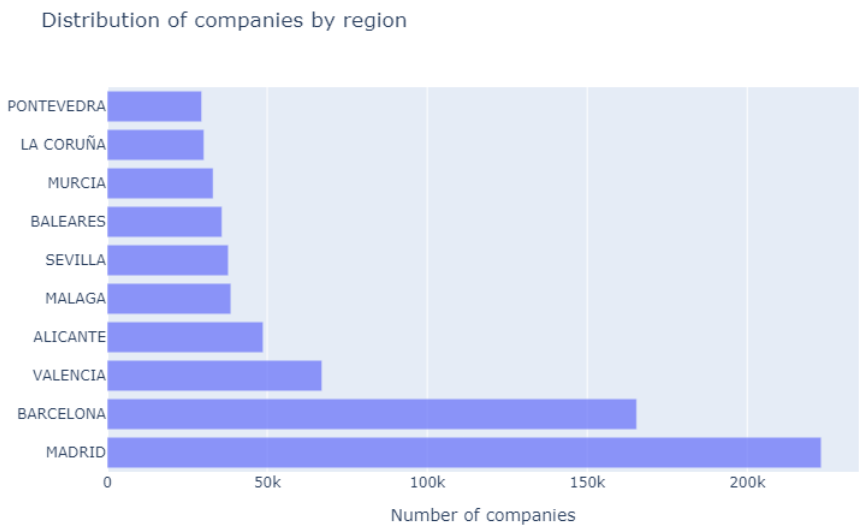

Figure 4: SME regional distribution

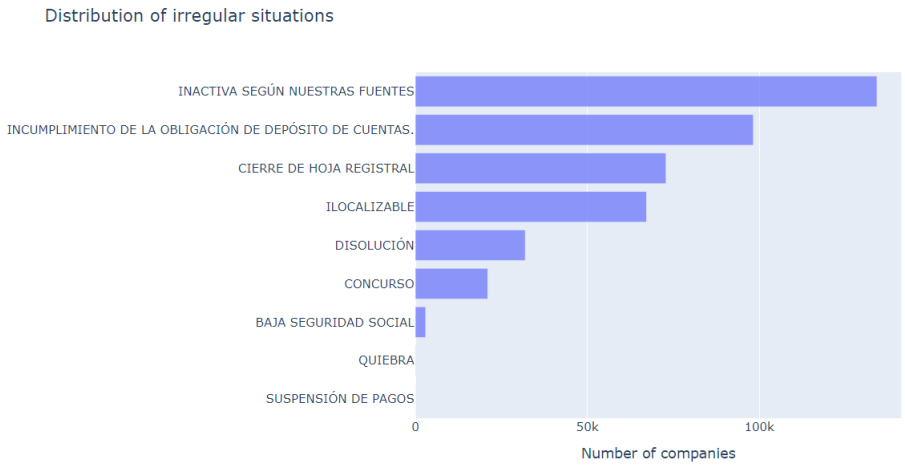

Figure 5: Irregular situations distribution

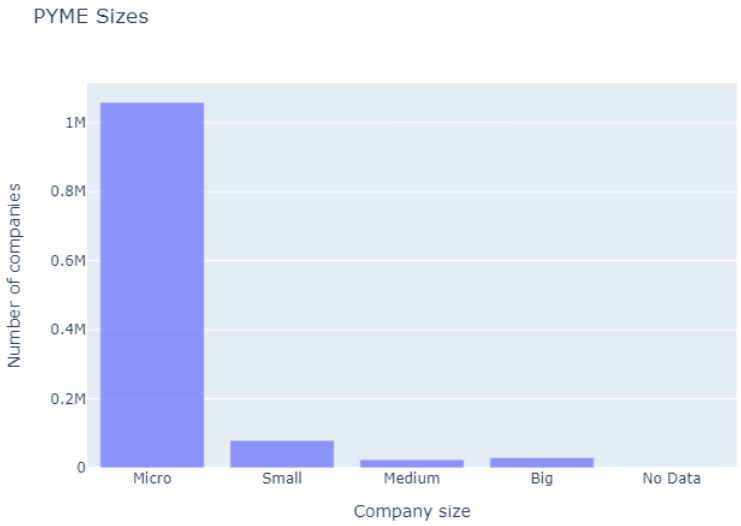

Figure 6: SME size distribution

3.2.2 Tables 2&3: Balance Sheet Headers & Detail

These two tables can be studied in tandem. The first table contains all the necessary information to allow for an easier organization and structuring of queries on the detail table. We can perform an exploratory data analysis on the former table. However, the latter table is too big to extract initial conclusions from. In

order to treat this data, several partitions of the table have been performed.

The BALANCE_SHEET_HEADERS table has one row for each year that each of the 1.2M companies has presented its financial statements. Thus, it has at most 1.2M x 14 rows (number of different companies x number of possible years). This table will be useful for knowing when an individual company has presented financial statements. Moreover, it provides overall information for the year such as NUM_EMPLOYEES.

The BALANCE_SHEET_DETAIL table has one row for each (CODE, VALUE) pair. Each company will have up to 88 different(CODE, VALUE) pairs per year. It may have more due to repeated values, multiple balances presented in one year and other inconsistency errors. A code is a number that identifies a single financial account(for example, accounts receivable). The value represents is a monetary amount. One example would be (12380, 20500) which would stand for 20,500€ in the Accounts Receivable account for company ID in year BALANCE_YEAR. This highly detailed approach means that this third table is massive. There are 352.8M (million) entries in this dataset. This is too big to handle by any regular computer or program so a special approach must be taken to read and treat this data. This process will be further detailed in the following section.

For now, some observations that can be extracted of the balance sheet headers table are shown in the items below. Supporting graphs for these conclusions are presented in figures 7-9.

- The number of companies in HEADERS is the same as in GENERAL_DATA, no additional considerations are needed in this regard.

- Each year has between 600k and 800k financial statement submissions by different companies. Out of all of these, 98% of companies present their annual results in December.

- Not all companies submit results every year and there are some accounts in their balance sheet which do not appear or are clearly incorrect (for example, negative values in asset accounts). Companies that present financial

results for more than seven years in a row, present a below-average (a-priory) probability of declaring bankruptcy.

- Most companies are classified as having a single size category over the 14-year time period. The more changes there are in this category, the higher probability of default within the company.

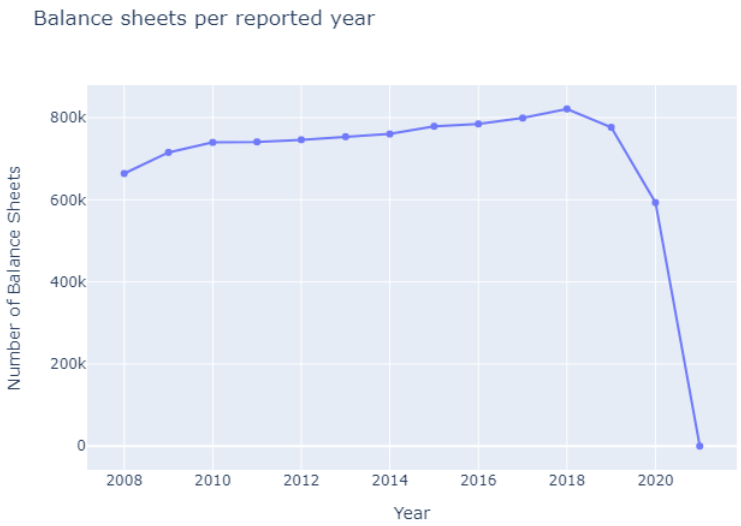

Figure 7: Yearly financial statement submissions

3.2.3 Table 4: Bankruptcy Data

This table has one row for each court or irregularity procedure initiated for a company at any point in time. Each instance indicates its source in the corresponding daily "BOE" publications and it has the code for the court in charge of the case. Most instances also have an "updated" date. These particularities will not be taken into account as a full analysis of this table would deserve an independent study. Thus, any company that appears in this table is considered to be in default. A few conclusions can be extracted from this data.

- There are 43k companies that appear at least once in this table. This represents 3.62% of the total number of available SMEs.

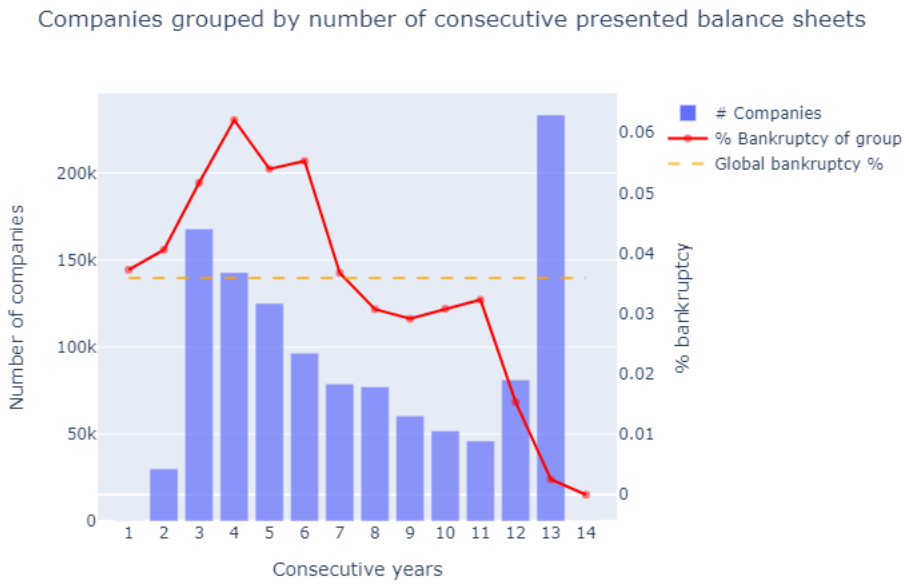

Figure 8: Distribution of number of consecutive years filed

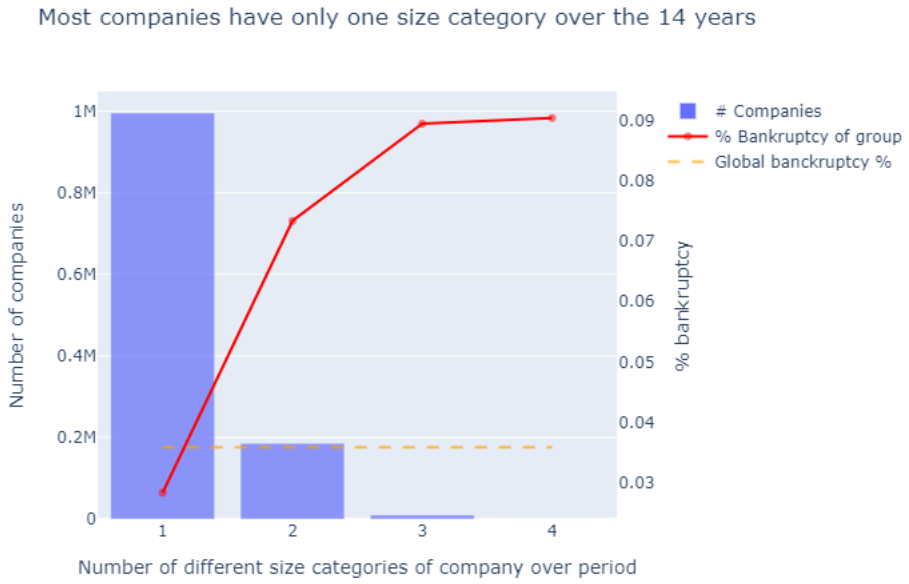

Figure 9: Distribution of size variations over time period

- A company appears in this table an average of 4.9 times. This implies that once a company faces any regulatory problems, it is highly likely that it will keep facing legal issues. Out of all companies entering any kind of legal status, less than 5% recover.
- Most procedures are updated. Only 485 out of a total of 213k entries do not present an "updated" status.

4 Data treatment

The data is treated in four different steps: table transformation and creation of yearly data sets, data addition and table merging, data cleaning and filtering, and ratio calculation.

The end goal is to arrive at an annual table that has one company per row and each of the columns represents a `CODE` for a balance sheet or P&L account (referred to only as balance sheet from now on for clarity and brevity purposes). Additional columns will then be added with the region, creation year, legal situation, and other information. We will have one of these tables per year. The following sections describe the process in detail.

All coding has been done with python notebooks and scripts and it is included in Oriol, 2022[3]. For further explanation regarding any of the steps described, the provided code may be consulted as it is easy to follow and it guides the reader through every step explaining all operations in detail.

4.1 Step 1: Obtaining the yearly data sets

In this step, information in `BALANCE_SHEET_DETAIL` is taken for each year and company and split up into different yearly data sets. We also add general company information available in `BALANCE_SHEET_HEADERS` such as `NUM_EMPLOYEES`.

First, `BALANCE_SHEET_DETAIL` is split up into 176 different files each containing approximately 20M entries. In each entry we apply different filters:

- Eliminate companies that have submitted less than 3 consecutive years of financial records (obtained in the previous step)
- Eliminate all 2021 balances due to having a relatively small sample size
- Eliminate duplicated balance sheet submissions. The latest one is not eliminated.

[3]Due to the length of the code, it has not been included within the current document. Code in github is set to private due to INFORMA terms. The author can be contacted for further details.

For each of the 176 files we transform the initial table into a "flattened version" with ID as rows and `CODE` accounts as columns. This is shown in figures 10 and 11.

	OID_EMPRESA	FECHA_CIERRE	CODIGO PARTIDA	VALOR
0	6278358	20201231	49500	13367,6
1	3990400	20201231	41500	-13130,66
2	36259305	20201231	21100	3000
3	36756968	20201231	12000	839,46
4	5987459	20191231	12382	8158,55

Figure 10: Initial partition of `BALANCE_SHEET_DETAIL`

CODIGO PARTIDA	10000	11000	11100	11200	11300	11400	11500	11600	11700	12000	...
OID_EMPRESA											
39	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
69	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
96	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
116	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
156	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...
206	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	259036.38	...

Figure 11: Flattened version of a `BALANCE_SHEET_DETAIL` partition

Once the flattening of the partitions is achieved, the same operation is performed over all 176 data frames. For each table, the rows corresponding to every year are extracted and combined with other rows of the same year. Thus, we transform 176 tables, each of which contains registers from 14 possible years into 14 tables - each of which contains all data for their corresponding year. One example is shown in figure 12. This constitutes a *yearly data set*.

OID_EMPRESA	10000	11000	11100	11200	11300	11400	11500	11600	11700	12000	...
5	70281.15	8210.88	NaN	-630.08	NaN	NaN	NaN	8840.96	NaN	62070.27	...
39	2621134.44	401887.11	2110.87	91878.34	273738.36	34159.54	NaN	NaN	NaN	2219247.33	...
41	2930382.0	1428672.0	NaN	1446.0	NaN	563712.0	838408.0	25106.0	NaN	1501710.0	...
46	131966.12	85248.74	NaN	85248.74	NaN	NaN	NaN	NaN	NaN	46717.38	...
47	360409.2	350247.3	NaN	308691.92	NaN	41555.38	NaN	NaN	NaN	10161.9	...
...	...	...	...	...	...	...	...	...	...	...	...

Figure 12: Example *yearly data set*

These yearly data sets are 14 tables. Each row represents a company that filed its financial statements at some point during the year. Each column is a financial account that the company can have submitted as part of its filing. As not all accounts are mandatory, some values may be NaNs.

4.2 Step 2: Including default data and number of workers

Once we have the yearly data sets described in the previous step, we must merge these tables with additional information contained in the tables `BALANCE_SHEET_HEADERS` and `BANKRUPTCY_DATA`. This is done via several merge operations controlling for the correct company and year. For example, a 2014 data set will only classify a company as bankrupt if its appearance in `BANKRUPTCY_DATA` is for 2014. Moreover, for a year t data set, additional columns are included for bankruptcies in years $t+1$ up to $t+4$ in case four year predictions want to be predicted. An example of the obtained tables is shown in figure13

4.3 Step 3: Adjusting accounts and adding additional enterprise information

This third step cleans the tables obtained in the previous steps and finalizes the adding of external information. To do this, several filters are applied to the resulting data frames and other data quality checks are performed. When they are finished, the final external information such as CNAE, company size (according to EU definition), and creation year is added.

OID_EMPRESA	10000	...	49300	NUM_EMPL_2008	N_DIFF_TAM	CONCURSO_2009	CONCURSO_2010	CONCURSO_2011	CONCURSO_2012	AÑO_CONCURSO	IS_BAD
5	80359.42	...	3346.56	6.0	1	0.0	0.0	0.0	0.0	0	0
39	7851763.48	...	-733811.12	23.0	1	0.0	0.0	0.0	0.0	0	0
41	2030989.0	...	25779.0	23.0	2	0.0	0.0	0.0	0.0	0	0
46	132814.95	...	-33370.66	4.0	1	0.0	0.0	0.0	0.0	0	0
47	353070.28	...	9055.16	0.0	1	0.0	0.0	0.0	0.0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
37242886	311167.58	...	-16979.66	2.0	1	0.0	0.0	0.0	0.0	0	0
37511421	911479.29	...	61596.71	5.0	1	0.0	0.0	0.0	0.0	0	0
37749826	1963332.99	...	9620.25	5.0	2	0.0	0.0	0.0	0.0	0	0
38113756	14094103.14	...	70825.37	28.0	1	0.0	0.0	0.0	0.0	0	0
38241573	7493593.7	...	-7702.9	18.0	3	0.0	0.0	0.0	0.0	0	0

Figure 13: Example merged yearly data set

We first detail the filters that have been applied to remove unwanted registers. The eliminated companies include those who filed invalid balance sheet versions, and companies with a fiscal year shorter than 12 months. These filters are detailed in table 4.

	2008	**2009**	**2010**	**2011**	**2012**	**2013**
Initial table size	642k	702k	732k	731k	736k	743k
Invalid financial template	0	4780	4847	4623	4523	4337
Invalid financial year duration	18.0k	14.8k	15.0k	15.4k	17.2k	20.5k
Invalid both template and dur.	0	38	14	9	15	18
Assets, Liab. and Equ. imbalance	2	3	16	42	17	0
Institutional companies	620	733	772	761	878	928
Final table size	624k	681k	712k	710k	713k	717k

Table 3: Detail of filters performed on final yearly data sets (I)

	2014	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Initial table size	750k	769k	774k	790k	809k	762k	587k
Invalid financial template	4195	4074	3920	3746	3498	3149	2473
Invalid financial year duration	18.4k	16.9k	20.9k	19.7k	17.4k	0.4k	0.3k
Invalid both template and dur.	14	20	10	15	14	6	4
Assets, Liab. and Equ. imbalance	0	0	0	0	0	0	0
Institutional companies	1008	1040	1111	867	790	740	521
Final table size	727k	747k	748k	766k	787k	758k	584k

Table 4: Detail of filters performed on final yearly data sets (II)

After the companies were filtered, the individual data for each register was examined. This analysis was performed to ensure that the ratios calculated in the

following section made sense. This is important because sometimes company filings are not done by experts, so some entries may be incorrect. Moreover, errors may occur when transferring the data from handwritten statements to computerized systems. Regarding the verification for the balance sheet and P&L accounts, the following rules were followed: If a balance sheet account has an incorrect sign, the sign is corrected (positive to negative). However, if the account is a P&L code, then we correct the sign and move the value to the correct P&L code. For example, if "Other revenues" is -20.000, then that account is set to 0 and 20.000 is added to the account "Other Expenses".

Finally, we add additional information such as CNAE, and company size according to EU definition; and we calculate different financial ratios that will be used in the final model. The result would be a final yearly data set like the one shown in figure 14

OID_EMPRESA	10000	40100	NUM_EMPL_2009	CONCURSO_2010	COD_SECTOR	LITERAL_CODIGO_PROVINCIA
5	70281.15	147233.76	4.0	0.0	I	ALICANTE
39	2621134.44	3450852.44	21.0	0.0	A	NAVARRA
41	2930382.00	1111730.00	28.0	0.0	B	BADAJOZ
46	131966.12	210794.52	4.0	0.0	I	BURGOS
47	360409.20	0.00	0.0	0.0	J	MURCIA
...	...	...	...	...	...	...

YEAR_CONSTITUCION	tamano_UE	R070	R071	IS_BAD	AÑO_CONCURSO	OID_EMPRESA
1982	Micro	0.064891	2.447229	0	0	**5**
1999	Pequena	0.000000	7.007774	0	0	**39**
1996	Pequena	0.000000	2.122380	0	0	**41**
1993	Micro	0.000000	0.802368	0	0	**46**
1997	Micro	NaN	0.000000	0	0	**47**
...	...	...	...	...	...	**...**

Figure 14: Example final yearly data set

4.4 Step 4: Data transformation integrity check and ratio calculation

The final step is to calculate the financial ratios that will be used for the different prediction models. However, before calculating the ratios, an integrity check is performed to ensure that the data transformation pipeline is working as intended. To perform the test, we will select several accounts for a random company from an original partition of `BALANCE_SHEET_DETAIL`. We will then check that those accounts correspond to the ones present in the final yearly data set. Figure 15 shows an example of a random extraction.

	OID_EMPRESA	FECHA_CIERRE	CODIGO PARTIDA	VALOR
0	6278358	20201231	49500	13367,6
1	3990400	20201231	41500	-13130,66
2	36259305	20201231	21100	3000
3	36756968	20201231	12000	839,46
4	5987459	20191231	12382	8158,55

Figure 15: Random extraction from `BALANCE_SHEET_DETAIL`

Company with OID 6278358 is chosen and we select a few other `CODE`, `VALUE` pairs from the same partition. The result is shown in figure 16. The color coding is explained in the following paragraph.

Figure 16 represents the raw data as has been handed to the author by INFORMA. We will now check that this data corresponds to the final values obtained in the yearly dats sets after all the transformations described above. To do this, we query several yearly data sets to obtain the corresponding value of the chosen accounts seen in the previous figure. For example, we will query the 2016 yearly data set for account 30000 of the corresponding company. We will do so with all four examples shown in table 16. The result of all these queries is shown in figure 17.

OID_EMPRESA	FECHA_CIERRE	CODIGO PARTIDA	VALOR
6278358	20161231	30000	1155019,75
6278358	20091231	32320	456372,66
6278358	20111231	41500	-8450,39
6278358	20171231	40100	44775

Figure 16: Additional `CODE`, `VALUE` pairs from the chosen company

OID_EMPRESA	ANO	30000	32320	41500	40100
6278358	2016	1155019.75	NaN	NaN	42433.33
6278358	2009	782110.35	456372.66	NaN	NaN
6278358	2011	1109832.33	NaN	-8450.39	42155.34
6278358	2017	1164697.29	NaN	NaN	44775.0

Figure 17: Data extracted from querying yearly data sets

The integrity check shows that it is highly unlikely that there is an error within the data processing pipeline. Thus, we can now compute all the financial ratios for each yearly data set. We have computed 162 financial ratios. Some of them have been calculated using two different versions of the ratio (some EBITDA definitions differ with regard to the accounts included). The definition and code for these calculations are included in Oriol, 2022.

5 Developing the prediction models

This section details the development of the bankruptcy prediction model. First, we test Altman's model with our current data to have a benchmark model against which our metrics can be tested. We also develop our own discriminant analysis model to try to improve Altman's predictions. Then, we develop classifiers with several techniques such as support vector machines, logistic regression, decision trees and neural networks. Out of the four experimental models, we choose the best alternative and try to improve it over several iterations. The resulting model will be the final classifier proposed by this study.

Before delving into the different classifiers, it is important to state several baseline considerations that are common for all developed models. First, as the data available is massive, we will randomly generate samples for the training of each model. These samples will be drawn from all yearly data sets (as we want our model to be able to generalize over several years) keeping the distribution of samples constant. This means that more negative (non-bankrupt) samples will be drawn whilst making sure that the proportion of positive samples stays the same as in the training data. Secondly, all data will be normalized as models have better performance with normalized data and some of our inputs may have significantly different magnitudes. Thirdly, the metric that will be used for model comparison is recall:

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

Recall is best used when the aim is to minimize the false negatives. It indicates the percentage of positive labels that have been correctly identified. In our case, we are focusing on correctly "catching" all failing companies at the expense of flagging some well-performing ones. In case we wanted to minimize the number of false positives, we would need to pay attention to the precision metric. This could be useful if we did not want any false positives to slip through. The formula for the precision metric is the accuracy of positive predictions:

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

A metric that combines these previous two is the f1 score. It strikes a balance between precision and recall. All models shown below will state all three of these metrics and the relative confusion matrices. This will allow the reader to understand the performance of each model and compare it with the others.

5.1 Discriminant analysis, Altman's model

First, we have applied Altman's model to the current data as a baseline. We have also used further iterations of the model with different weights proposed by Altman over the years as an improvement to his own model[4]. Finally, we have trained a discriminant model ourselves to see if this simple approach will be sufficient. Thus, we have four different results and confusion matrices.

	Classic Altman	**Altman v1**	**Altman v2**	**Trained DA**
precision	0.17	0.19	0.22	0.52
recall	0.87	0.84	0.55	0.07
f1-score	0.29	0.31	0.31	0.13

Table 5: Discriminant analysis — Metrics

		Predicted Label				
	Classic	Non bankrupt	Bankrupt	**Altm. v1**	Non bankrupt	Bankrupt
	Non bkrt.	22%	63%	Non bkrt	29%	56%
True Label	Bankrupt	1.9%	13.1%	Bankrupt	2.4%	12.6%
	Altm. v2	Non bankrupt	Bankrupt	**Trained DA**	Non bankrupt	Banktupt
	Non bkrt	55%	30%	Non bkrt	84%	1%
	Bankrupt	6.8%	8.2%	Bankrupt	13.9%	1.1%

Table 6: Discriminant analysis — Confusion matrices

We can extract two main conclusions from the previous data. The first one is that Altman's models have a high recall metric even when applied to this new

[4]For the purposes of the metrics and confusion matrices, companies in the "grey area" of Altman's model have not been classified as bankrupt.

set of data. However, the f1 score and overall performance are rather poor as many firms are incorrectly classified as bankrupt. The second conclusion is that Altman seemed to prioritize a higher recall when choosing the weights for its models, even at the expense of global model performance. This still holds true with SME data. Moreover, our trained discriminant analysis performed incredibly poorly. The explanation for this is that additional variables have been used for the discriminant analysis other than Altman's five ratios. This introduction of new variables (further detailed in the following section) is because using only Altman's ratios for model retraining causes recall to be 1 at the expense of classifying almost all samples as bankrupt. According to the previous considerations, The best model of the four shown here would be Altman v1. Even thou

5.2 Four experimental models: SVM, Log Reg, Decision Tree, Neural Network

In this section, we will try out four different classification models to see if we can obtain better results than the previous discriminant analysis. In order to achieve this, new variables have been introduced to complement the ones used by Altman in his models. Historically, he only used the ratios explained in section 2. However, we have used new variables that improved the performance of these experimental models such as CNAE code, and company size and age. Tables 7 and 8 show the results for these models arranged in a similar manner as the discriminant analysis models. The Relative confusion matrices and the other obtained metrics allow for a fair comparison between models.

	Cl. Altman	**SVM**	**Log. Reg.**	**Dec. Tree**	**Neural Ntwk.**
precision	0.17	0.22	0.23	0.5	0.23
recall	0.87	0.76	0.67	0.09	0.73
f1-score	0.29	0.35	0.34	0.15	0.35

Table 7: Experimental models — Metrics

As a general conclusion, we can say that the results are not too promising either. However, performance seems to be better than the previous section's models (not

		Predicted Label				
True Label	**SVM**	Non bankrupt	Bankrupt	**Log. Reg**	Non bankrupt	Bankrupt
	Non bkrt.	46%	39%	Non bkrt.	52%	33%
	Bankrupt	3.5%	11.5%	Bankrupt	4.9%	10.1%
	Dec. Tree	Non bankrupt	Bankrupt	**N. Ntwk.**	Non bankrupt	Bankrupt
	Non bkrt.	83.7%	1.3%	Non bkrt.	49%	36%
	Bankrupt	13.7%	1.3%	Bankrupt	4%	11%

Table 8: Experimental models — Confusion matrices

for the decision tree). Even though recall is somewhat lower, there is less tendency to incorrectly classify a company as bankrupt which was an important flaw of Altman's application. Overall, we can say that the best performing models are the support vector machine and the neural network. We have chosen to further delve into the latter one. The reason for choosing the neural network is because we can have some visualization of the training process and understand whether it is improving as time passes or not. Moreover, we can further adjust its architecture and try to improve the results. Whilst it is true that we could try using other kernels with the SVM, several have been tested and the results have not been better. Thus, we have chosen to fine-tune the neural network.

Before trying to improve it by adding different layers and better overfitting controls, we will show the obtained neural network and its training process.

Figure 18 displays the neural network's architecture. It has 28 different entries (some variables such as the CNAE group were one-hot encoded) in the input layer. This input layer is followed by three hidden layers of 20, 10, and 5 nodes respectively. This results in a total of 851 trainable parameters (weights). All activation functions are *relu* except for the *sigmoid* activation function in the last layer. To control overfitting, dropout layers are located between each dense layer. The loss function used for the backpropagation is binary cross-entropy with an adam optimizer. As a final note, errors made with "bankrupt samples" are given much more importance (weight) than misclassifications in non-bankrupt entries.

The network's training performance is indicated in Figure 19 which may require some additional explanation to interpret. The vertical axis shows loss (training and validation). Greater loss means more prediction error. Training stands for data

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	[(None, 28)]	0
dense_37 (Dense)	(None, 20)	580
dropout_12 (Dropout)	(None, 20)	0
dense_38 (Dense)	(None, 10)	210
dropout_13 (Dropout)	(None, 10)	0
dense_39 (Dense)	(None, 5)	55
dropout_14 (Dropout)	(None, 5)	0
dense_40 (Dense)	(None, 1)	6

Total params: 851
Trainable params: 851
Non-trainable params: 0

Figure 18: Initial neural network layers and architecture

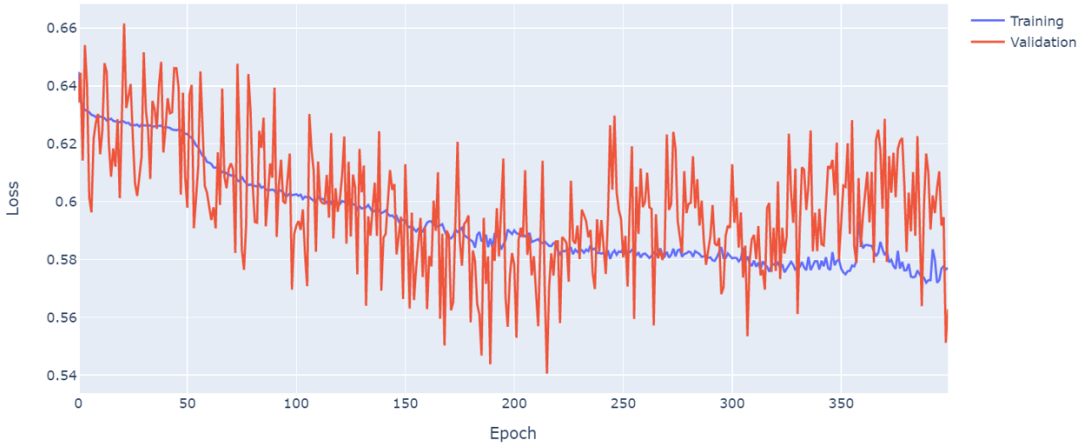

Figure 19: Initial neural network training process

used to build the model and validation is the data used to test the model. This second set of data is different to ensure the the model is useful for data other than the one used in training (reducing model variance). The horizontal axis indicates the epoch. Training a neural network consists using all training data as many

times as the number of epochs indicates. The model's weights are updated every batch. In this case, data was passed through the network 400 epochs with a batch size of 500 samples. This means that weights were updated every 500 samples and all samples go through the model 400 times.

Loss stays fairly constant throughout the whole training process and almost no improvement is made. The validation error is more variable but fairly constant throughout as well. Overall, some overfitting can be observed around the epoch 250 where validation error stops decreasing steadily and stays constant whilst training error keeps decreasing.

Section 5.3 will discuss the two alternative methods that have been tested to improve these results: a slightly more complex neural network and an autoencoder.

5.3 Improving the neural network

Two possible improvements have been tested. The first one is a variation of the previous neural network's architecture, which tries to capture a higher amount of information from the provided data and has a higher number of trainable parameters. The second one is an autoencoder. Autoencoders are a kind of neural network architecture that is highly useful in detecting outliers and we have tested this approach with the hypothesis that future bankrupt companies will be considered "outliers" by the neural network.

5.3.1 Neural network improvements: changing the architecture

The first method tested is changing the number of layers, nodes and the overall network layout. Figure 20 shows the new architecture.In this case, it has four hidden dense layers with a higher number of nodes in each layer (100, 50, 25, 10). Dropout layers are introduced in between each dense layer to control overfitting. This results in a total of almost 9,500 trainable parameters. All in all, the network reports some improved results over the previous models. However, performance is still not too promising.

For comparison purposes, we show this network's training performance in Figure 21. For this case, we used the same number of epochs and samples per batch

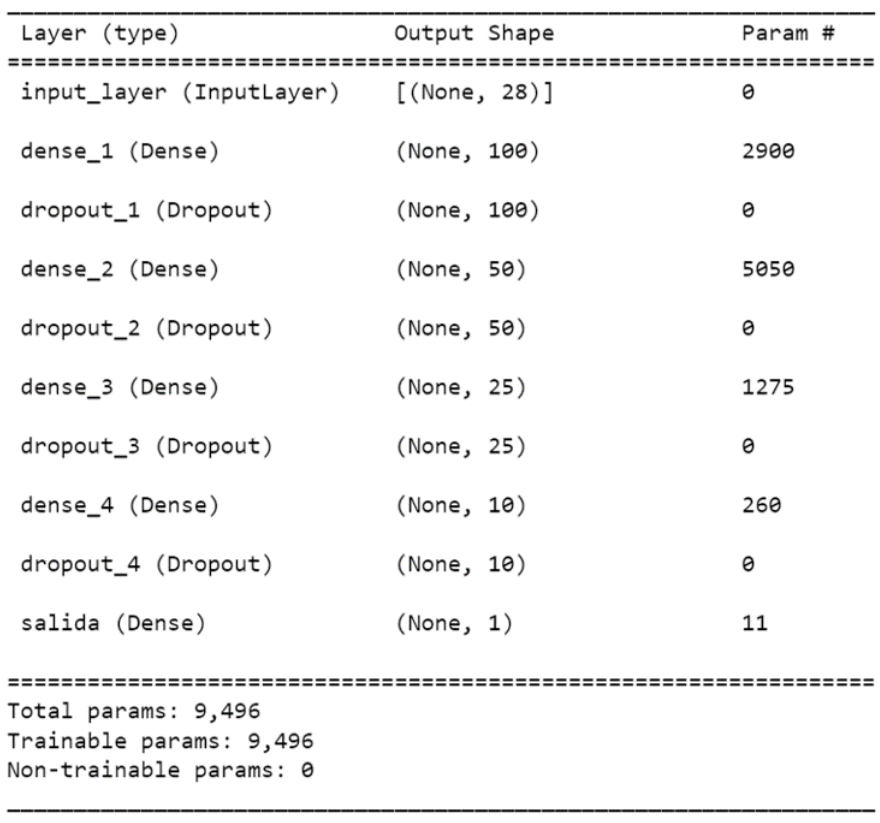

Figure 20: Improved neural network layers and architecture

as in the previous one. The network is more complex than the previous one, thus it would make sense to train with a higher number of epochs. However, we saw in the previous case that the model started overfitting around epoch 250. For this reason, 400 epochs is used as a valid number.

The results are slightly more promising as the loss is lower in this more complex case. However, overfitting can also be observed passed epoch 250. The confusion matrix and other metrics are compared with the autoencoder in section 5.3.3 and are - to some degree - better than the experimental models shown above.

5.3.2 Neural network improvements: autoencoder

Autoencoders consist of two distinct halves: an encoder and a decoder. Encoders reduce entry samples to a lower number of dimensions, whereas decoders recon-

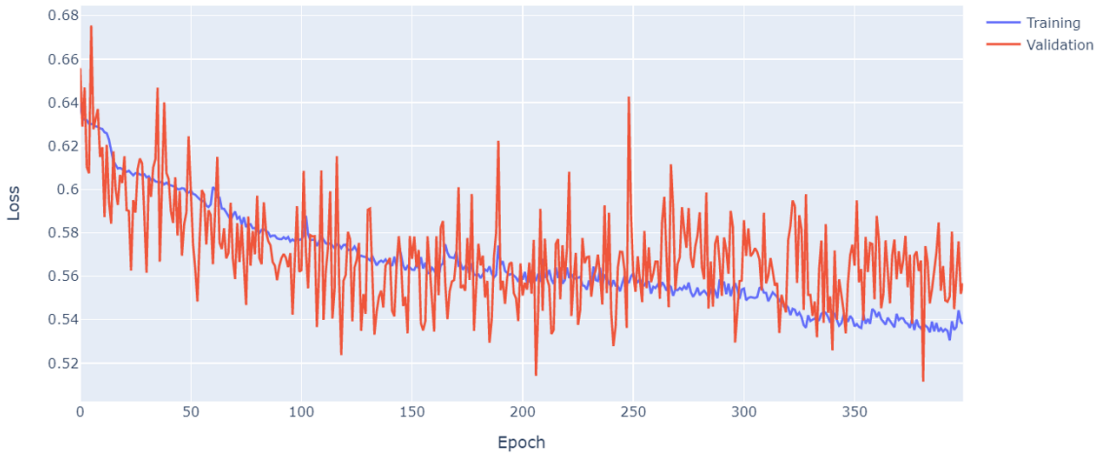

Figure 21: Improved neural network training process

struct samples from a low number of dimensions. Thus, an autoencoder - by combining these two neural networks - is effectively deconstructing and reconstructing all samples. The training process does not take into account the labels but strives to learn how to rebuild negative samples with the minimum possible error. Theoretically, the resulting model will output higher errors when positive (bankrupt) samples are introduced. By analyzing this "reconstruction error", and setting the adequate threshold we will be able to discriminate between the two kinds of companies.

Figure 22 shows the autoencoder architecture at from a higher perspective. It consists of both the encoder and decoder, with the same entry and exit sizes. Overfitting is controlled by l1 regularization layers. Figures 23 and 24 show a more granular detail of the encoding and decoding layers and the trainable parameters in each one of them.

Once a sample goes through the autoencoder, an error metric can be constructed by comparing the entry data with the output. In our case, MSE was the chosen metric for this comparison. In theory, positive samples will present a different error profile than negative samples. The reasoning behind this is that the model will learn how to deconstruct and reconstruct negative samples. Thus, when a positive entry is introduced a higher reconstruction error will show that it

```
Model: "SparseAutoencoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 InputSparseAutoencoder (Inp  [(None, 28)]             0
 utLayer)

 SparseEncoder (Functional)  (None, 8)                 1345

 SparseDecoder (Functional)  (None, 28)                1365

=================================================================
Total params: 2,710
Trainable params: 2,710
Non-trainable params: 0
_________________________________________________________________
```

Figure 22: Sparce autoencoder layers

```
Model: "SparseEncoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 InputEncoder (InputLayer)   [(None, 28)]              0

 DenseEncoder1 (Dense)       (None, 25)                725

 DropoutEncoder1 (Dropout)   (None, 25)                0

 DenseEncoder2 (Dense)       (None, 18)                468

 DropoutEncoder2 (Dropout)   (None, 18)                0

 OutputEncoder (Dense)       (None, 8)                 152

=================================================================
Total params: 1,345
Trainable params: 1,345
Non-trainable params: 0
_________________________________________________________________
```

Figure 23: Sparce encoder layers

is different to the rest. The distribution of this reconstruction error is shown in two graphs. The first graph, in figure 25 is a histogram of the relative frequency of the reconstruction error for both kinds of companies. It can be seen, however, that even when training only with negative samples, the error distribution does not seem to differ significantly between classes. The second plot (figure 26), is a box-plot showing the training reconstruction error for default and non-default cases and the overall dataset. In this case, it is even more surprising to see that the default distribution has lower values than its opposing one. Nevertheless, the distributions do not differ significantly and the most relevant conclusion to be

```
Model: "SparseDecoder"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 InputDecoder (InputLayer)   [(None, 8)]               0

 DenseDecoder1 (Dense)       (None, 18)                162

 DropoutDecoder1 (Dropout)   (None, 18)                0

 DenseDecoder2 (Dense)       (None, 25)                475

 DropoutDecoder2 (Dropout)   (None, 25)                0

 OutputDecoder (Dense)       (None, 28)                728

=================================================================
Total params: 1,365
Trainable params: 1,365
Non-trainable params: 0
_________________________________________________________________
```

Figure 24: Sparce decoder layers

extracted from these graphs is that both kinds of companies rarely differ on the given metrics.

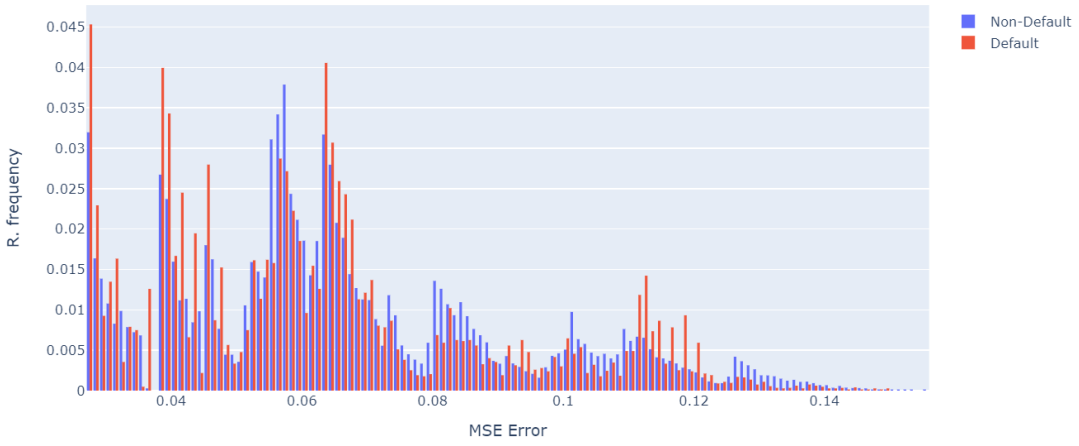

Figure 25: Reconstruction error histogram

Once, we have the error metric computed for the training cases, a threshold has to be chosen to discriminate between bankrupt and non-bankrupt. In the histogram shown in figure 25, this would be represented as a vertical line at any MSE value. Any sample to the left of this threshold will be classified as negative (non-bankrupt) and samples to the right - with higher reconstruction errors -

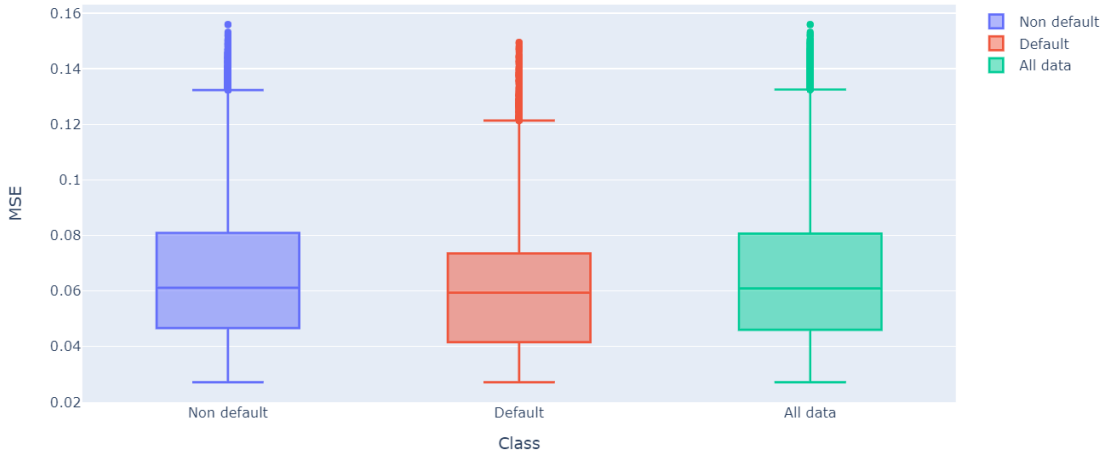

Figure 26: Reconstruction error distribution

will be labelled as positive. To decide the value of this threshold metric, several possible values are tested and our performance metrics are plotted against them. Effectively, we have multiple different classification models, as depending on the threshold value, a different confusion matrix will be obtained. In this case, a within the (0.035-0.045) range will be chosen as we want a high recall without losing too much performance in the f1 score metric. The confusion matrix obtained by this classifier is compared against improved neural network's one and it is shown in section 5.3.3.

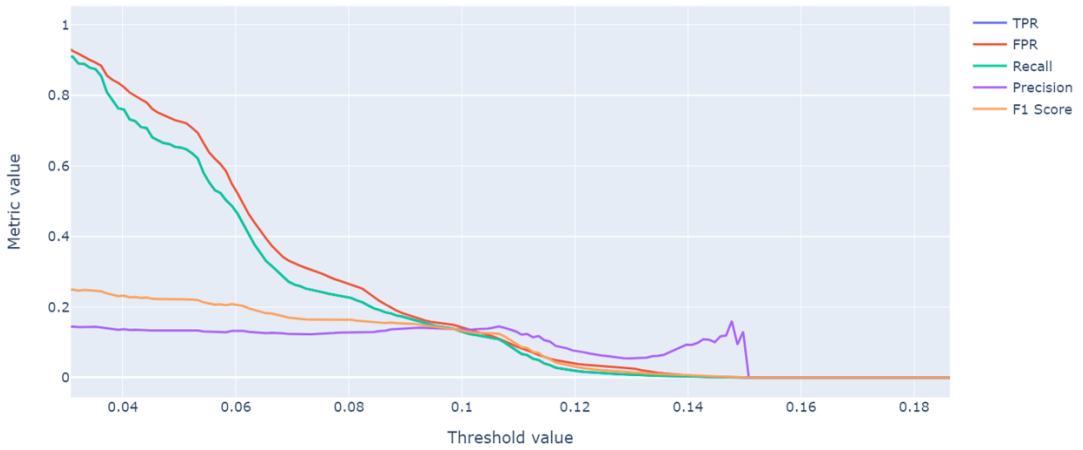

Figure 27: Threshold evaluation

An intuitive way to understand this graph is that the further right the threshold is chosen along the x-axis, the more "strict" the model is when deciding whether a company is classified as bankrupt or not. Therefore, higher thresholds, cause samples in the confusion matrix to shift from the right column to the left column. The ideal model would have samples in the top row (incorrectly labelled as positive) switch faster than those in the lower row (correctly labelled positives). This is the balance that we aim to strike by balancing the recall and precision metrics.

5.3.3 Neural network improvements: performance results

With the two new models constructed, we can show their respective metrics and the confusion matrices for each of them. Tables 9 and 10 show these results.

	Cl. Altman	**N. Ntwk**	**Improved N. Ntwk**	**Autoencoder**
precision	0.17	0.23	0.24	0.15
recall	0.87	0.73	0.77	0.86
f1-score	0.29	0.35	0.37	0.25

Table 9: Improved models — Metrics

		Predicted Label				
True Label	**Imp. Ntkw**	Non bankrupt	Bankrupt	**Autoenc.**	Non bankrupt	Bankrupt
	Non bkrt.	49.3%	35.7%	Non bkrt.	9.7%	75.3%
	Bankrupt	3.4%	11.6%	Bankrupt	2.1%	12.9%

Table 10: Improved models — Confusion matrices

As expected from the graphs represented in the previous section, the autoencoder's performance - whilst better than some of the experimental models - is rather poor as well. However, the improved neural network shows promising results with high recall values and a higher f1-score than all other models. By the indications of the confusion matrix, a "non-bankrupt" prediction would be correct more than nine times out of ten ($1 - \frac{3.4}{49.3} = 93.1\%$). Even though this was not the desired result - we wanted to be correct on "default" predictions - it also has its purposes. This is further explained in section 6.

All in all, we can safely say that the improved neural network does improve our baseline model considerably and provides more accurate predictions for our

use case. However, it should be noted that with the current dataset and method of study it seems that these models would require additional information or a separate parallel approach to be fully effective as overall performance is not great.

6 Conclusions

This study presents three main conclusions. The first is a general approach and code (provided in Oriol, 2022) to gathering, cleaning and organizing in clearly readable tables financial data regarding Spanish companies. The second conclusion is the results of applying Altman's model and its variations to a more updated economical environment and a specific set of companies. Finally, we have provided a model that slightly improves Altman's performance when predicting the default of SMEs based only on public data.

The provision of a generalized method to regularize financial data is not a trivial matter. With the code provided, any financial data provided by Spanish companies (as they use similar accounting models with the same overall data structure) can be cleaned and organized in annual data sets. This can be the stepping stone for further studies and papers finding relationships between companies and comparing them across different categories. Some improvement that could be made to the provided code is parallelization of some of the programming tasks, especially in the analysis of `BALANCE_SHEET_DETAIL`.

Altman's model - and especially its variations - has been successfully applied to the current data. A clear conclusion that can be extracted based on the results, is that Altman also seemed to prioritize recall to precision. This validates our approach and justifies the parameter and threshold decisions taken in the optimization of all of the provided models. Altman's model still works, albeit weakly and classifying too many companies as bankrupt.

Finally, we have provided an alternative approach that improves Altman's baseline when applied to Spanish SMEs. Even though the model is not perfect, it provides a certain level of security when classifying a company as non-bankrupt. This was not the initial objective, however, it is a useful byproduct of the study. A disadvantage of this model is that it is less understandable than Altman's discriminant analysis due to the opaque nature of neural networks.

For further improvement and future work, two options are proposed. First, using some private data, from the perspective of either the SME or the financial

entity servicing it would probably improve the model. The combination of this data with some of the models presented would most likely improve the results. Secondly, having time-series data would allow for some historical metrics of a business to be evaluated - not just a snapshot in time, but the evolution and tendency of its accounts. This would probably result in remarkable improvement. This approach was not taken in this study due to having at most 14 data points for each business. However, it could be undertaken if we had monthly data for example.

In conclusion, the obtained model moderately achieves the initial objective of improving and understanding the baseline. However, it fails to be good enough to be of any practical use. Nonetheless, this study and the data organization may serve as a valid stepping stone for developing future models and other studies regarding SME financial statements and operating situations.

References

Altman, E. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy". In: *The Journal of Finance* 23.4.

Altman, E. et al. (2017). "Financial Distress Prediction in an International Conext: A Review and Empirical Analysis of Altman's Z-Score Model". In: *Journal of International Financial Management & Accounting* 28.2.

Blanco, R. et al. (2021). "Impact of the COVID-19 crisis on Spanish firms' financial vulnerability". In: *Banco de España* 2119.

Blanco Ramos, F., M. Fernández Blanco, and M. Ferrando Bolado (2016). "El impacto de la crisis económica en las PYMEs Españolas". In: *Colegio de Economistas de Madrid* 149, pp. 66–79.

Camacho-Miñano, M., M. Segovia-Vargas, and David Pascual-Ezama (2013). "Which Characteristics Predict the Survival of Insolvent Firms? An SME Reorganization Prediction Model". In: *Journal of Small Business Management*.

Cifras PYME Enero 2022 (2022). Dirección General de Industria y de la Pequeña y Mediana Empresa. Ministerio de Industria, Comercio y Turismo.

El número de pymes que se declararon en quiebra en España se ha triplicado desde el comienzo de la crisis (2022). Europa Press. URL: https://www.europapress.es/economia/noticia-economia-numero-pymes-declararon-quiebra-espana-triplicado-comienzo-crisis-20131126163223.html (visited on 03/10/2022).

GDP growth, annual % (2022). World Bank. URL: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2020&locations=ES-EU-1W&start=2007 (visited on 03/10/2022).

Guest, N. (2021). *Financial Statement Analysis, Class Notes*. Cornell SC Johnson College of Business.

Guía del Usuario sobre la Definición del Concepto de PYME (2016). 956541. Comisión Europea. Mercado Interior, Industria, Emprendimiento y PYMES.

Hernández, A. and MP. Buil Vilalata (2012). "SME financing in Spain: short-term emergency, long-term challenge". In: *CaixaBank Research*.

INFORMA Filial de Cesce líder en el suministro de Información Comercial, Financiera, Sectorial y de Marketing de empresas y empresarios.

Keasney, K., J. Pindado, and L. Rodrigues (2015). "The determinants of the costs of financial distress in SMEs". In: *International Small Business Journal* 33.8, pp. 862–881.

La PYME Española y el reto del crecimiento (2022). El Confidencial. URL: `https://datos.elconfidencial.com/sage-6/` (visited on 02/17/2022).

Malakauskas, A. and A. Lakstutiene (2021). "Financial Distress Prediction for Small and Medium Enterprises Using Machine Learning Techniques". In: *Inzinerine Ekonomika-Engineering Economics* 32.1.

Marco Estratégico en Política de PYME 2030 (2021). Colección Panorama PYME. Ministerio de Industria Comercio y Turismo.

Oriol, N. (2022). *Code for all data processing and model creation.* URL: `https://github.com/noriolg/pyme-default-prediction`.

Retrato de la PYME (2022). Dirección General de Industria y de la Pequeña y Mediana Empresa. Ministerio de Industria, Comercio y Turismo.

SME Definition (2022). European Commission. URL: `https://ec.europa.eu/growth/smes/sme-definition_en` (visited on 02/17/2022).