RISK OF BUSINESS FAILURE

STUDENT GROUP

Jose Gonzalez,
Ivan Ogando,
Andres Cordeiro

Statistics for Data Science

Data Science & Business Informatics

Academic Year 2021/2022

Contents

1 Introduction

For this project the motivation was the investigation of the risk factors of business failure of a set of Companies from a given dataset. The investigation was segmented into 5 questions that were proposed by the project. The questions from A to C ask for study over the distribution of ceirtan aspects of the dataset. Questions D and E focus more on the answers given by using parametric models for failure scoring and comparing the scoring with a given technique.

1.1 The Dataset

The dataset given was composed of 1,894,412 entries and 80 total columns which contain from the name, ATECO classification, legal status and incorporation year of the company to performance metrics from the last 3 registered years. There are a total of 7 variables without missing values. 18 categories for *Legal Form*, 10 categories of *Legal Status*, and 20 recorded *Regions*.

1.2 Data Preparation

Given the proposed questions for this project we needed to create some variables that allowed us to work better with the information given. We needed to establish an age and a size for each company. We also understood that the *ATECO 2007 code* could be helpful but not as it was as well as the *Legal Status* since it was our main classifier however the 10 categories could be simplified.

We opted to create 4 new features:

1. *Age* :Describing the years the company has/was active. This was created by the difference of the "Incorporation year" and the last "Last accounting closing date".

2. *Size* : Segmenting the companies into the following matrix according to the Small and medium-sized enterprises of the European Union:

   - *Micro* : Less than 10 workers and a balance sheet total less or equal than 2 millions
   - *Small* : Less than 50 workers and a balance sheet total less or equal than 10 millions
   - *Medium* : Less than 250 workers and a balance sheet total less or equal than 43 millions
   - *Large* : We classified as large any of the companies that went beyond the medium limits.

3. *ATECO* : Taking the codes of the *ATECO 2007 code* column we assigned to every company their respective sector on the most general level. All according to the segmentation available in the "Instituto Nazionale di Statistica (Istat)".

4. *Status* : Simplifying the *Legal Status* as Active or failed. For the interest of this investigation the 3 active tags were merged into one simple "Active" and the rest were considered "Failed"

After the creation of these attributes we preceded to create 2 datasets to work with. One for the questions A, B and C and another for the rest. For the first dataset since the questions required to analyze the different distributions according to age and size and observe the change for company sector and legal form, we opted to select all the new features and include the region and Last year. For the second dataset, to be used in question D and E, we opted for all the

financial indicators of the last accounting year of each company, plus the new Status and Age features. Additional variables and transformation were required for these last 2 questions, but these are explained in section 5.

2 Comparison for the distributions of size/age/other between failed and active companies at a specific year

In this section we analyzed the densities of the companies comparing the failed with the active sections by the year 2019. We aimed to investigate if they changed for a given ATECO sector, Legal form or Region. We utilized the Kolmogorov-Smirnov test (KS test) to compare the distributions since we could visualize that we needed a non parametric test for these cases. We selected the year 2018 since it was the year with the most records.

2.1 Analysis by age

2.1.1 Legal Form

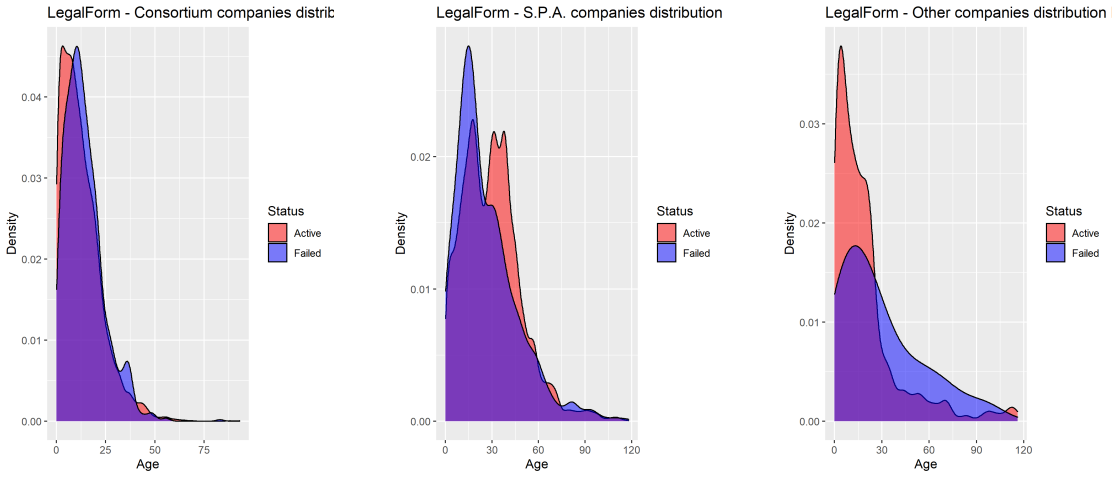


Figure 1: Active failed Densities by Legal Form

As we can see from the most interesting cases on Figure 1 most of the companies in failed and active follow a similar distribution. However, there was a notable peak in the earlier years for the failed companies for 4 of the stated cases. The S.R.L. companies have presented the most unique distribution since it consisted of several peaks however both distributions of the failed and active seemed to overlap at each point as in the S.N.C distribution. We only selected the 3 most representative out of the 9 plots, the rest can be seen in the last section.

In table 1 we could observe that the greatest distance was present in the sectors "Other", "S.N.C." and "Social cooperative". Figure 2 confirmed the results obtained on the test, showing how *Consortium* one of the lowest D values distributions almost match also, it was observable how *Other* the highest of the results had very different distributions. In *Other* was possible to view how there was a high peak from active companies pertaining to the 0 to 30 years however the descent was very notable as they approached 15 years and 25-29 years. The failed companies distribution from *Other* peaked near the 15-20 years zone however it did not present the steep slope from its counterpart.

Sector / Metric	D	P-value
Consortium	0.10031	1.e-07
Other	0.2749	0.1228
S.A.S.	0.1586	0.5983
S.C.A.R.L.	0.08529	4.e-09
S.C.A.R.L.P.A.	0.0496	1.e-08
S.N.C.	0.1884	0.7664
S.P.A.	0.1412	0
S.R.L. one person	0.0566	0
S.R.L. simplified	0.0745	0
S.R.L.	0.07510	0
Social cooperative	0.1445	5.e-06

Table 1: Results for KS test for question A according to age by Legal Form

We did not emphasise the results for the p-value since they were not the most reliable. This was attributed to there being a large amount of data. The p-value tends to find trivial deviations from the null hypothesis the larger the amount of data it works with. We presented non specific p-values that were extremely low values as 0 on the tables for this and the next sections.

2.1.2 ATECO

Metric	A	B	C	D	E	F	G	H	I	J	K
D	0.0856	0.1246	0.0406	0.0844	0.0855	0.1655	0.0607	0.0699	0.0798	0.0605	0.1434
P-value	3.e-07	0.0332	3.e-13	0.0004	0.0112	0	0	7.e-10	0	8.e-13	0

Metric	M	N	O	P	Q	R	S	T	U	No Sector
D	0.0921	0.0721	0.2285	0.0554	0.0915	0.1105	0.0795	0	0.6536	0.05336
P-value	0	0	0.9207	0.0479	2.e-09	1.e-15	1.e-05	0	0.0143	0.3763

Table 2: Results for KS test for question A according to age by Sector

In the table 2 we could observe how there were several cases with high D values. However, none as great as the *U-ORGANIZZAZIONI ED ORGANISMI EXTRATERRITORIALI* that presented a 0.65 as it could be seen in Figure 2. The fail distribution for this sector had 3 notable peaks however, the highest was covering companies from 35 to 40 years.This peak was almost completely separated from the peak of the Active companies that was near the 20 years of age.

An anomaly that was presented was the sector *T-ATTIVITÀ DI FAMIGLIE E CONVIVENZE COME DATORI DI LAVORO PER PERSONALE DOMESTICO* for the selected year the dataset did not present any failed records. This was an issue that we had to work around for the test. This was the reason the results were denoted with NAN. We understood that a wider span of year could probably solve the issue.

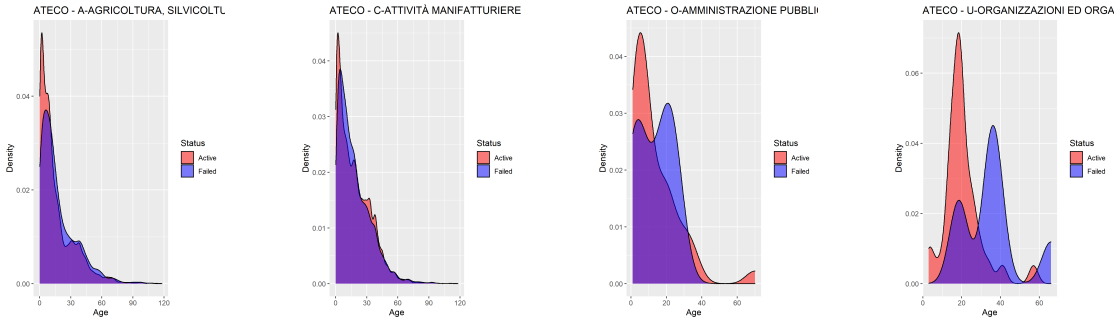


Figure 2: Active failed Densities by ATECO

2.1.3 Region

For the region distributions we found that most had similar distributions. A pattern appeared from the lowest D value to the highest. Most of the Failed and Active distributions tend to be high in the young companies zone. Mainly around the 13-15 year mark. In figure 3 we saw how the shapes of the distribution for Basilicata, Calabria and Molise follow the same pattern however their peak was not so pronounced as the one for active companies. This was what made them the ones with the highest D values. We note that even thought those were the highest values they are very low in comparison to the ATECO and the Legal Form segmentation.

In the graphs with lowest D values from the test we note that they also have an overlap of almost 20 years from the oldest companies. The ones with the most distance had an oldest company near the 100 years. The ones with least distance had their oldest companies near the 120 years. This was just a detail we noticed but they are not the most relevant since their distribution was from 60 years onward.

Metric	Abruzzo	Basilicata	Calabria	Campania	Emilia R.	Friuli	Lazio	Liguria	Lombardia	Marche
D	0.1052	0.1363	0.1761	0.1354	0.0542	0.0438	0.0837	0.07	0.0433	0.0931
P-value	2.e-13	6.e-07	0	0	2.e-16	0.0192	0	1.e-07	0	0

Metric	Molise	Piemonte	Puglia	Sardegna	Sicilia	Toscana	Trentino	Umbria	Valle	Veneto
D	0.1373	0.0597	0.1145	0.1146	0.1585	0.0691	0.0848	0.0831	0.0898	0.0523
P-value	0.0002	3.e-11	0	1.e-13	0	0	5.e-07	4.e-06	0.3253	1.e-15

Table 3: Results for KS test for question A according to age by Region

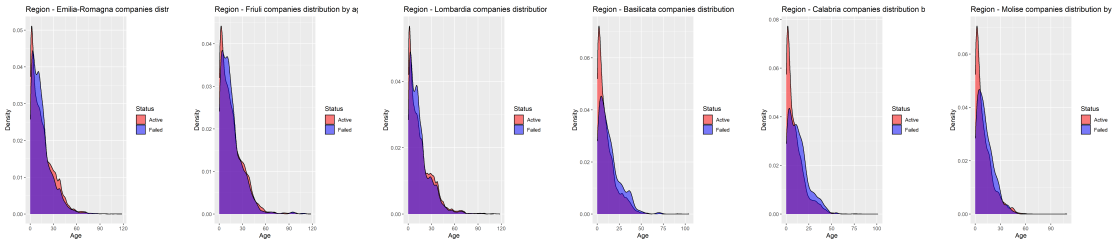


Figure 3: Active failed Densities by Region

2.2 Analysis by Size

For this section we analyse the same sections but dividing them in the size of the company. As stated before, we utilized the same parameters as the European Union standards. The micro companies dominated the dataset, for this reason we went with a logarithmic scale so we could appreciate the distributions details better.

2.2.1 Legal Form

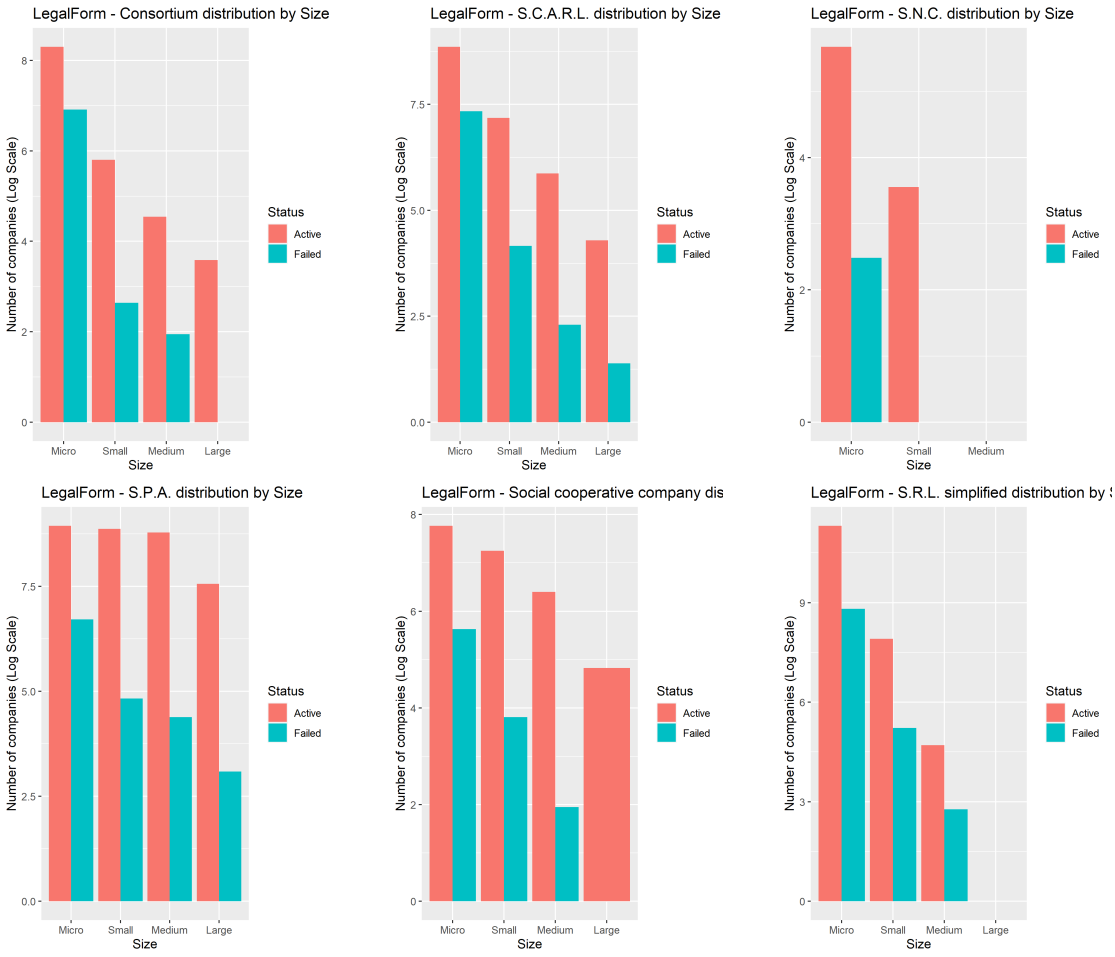


Figure 4: Active failed Densities by Size and Legal Form

As exhibited on the Figure 4 the majority of the distributions tend to have more micro companies. Consortium presents very few large companies and mostly as active, the S.N.C had very few records itself in comparison to the others however it still followed the tendency of having more micro companies than any other category. Social cooperative did not present any large companies failed records for this year.

2.2.2 ATECO

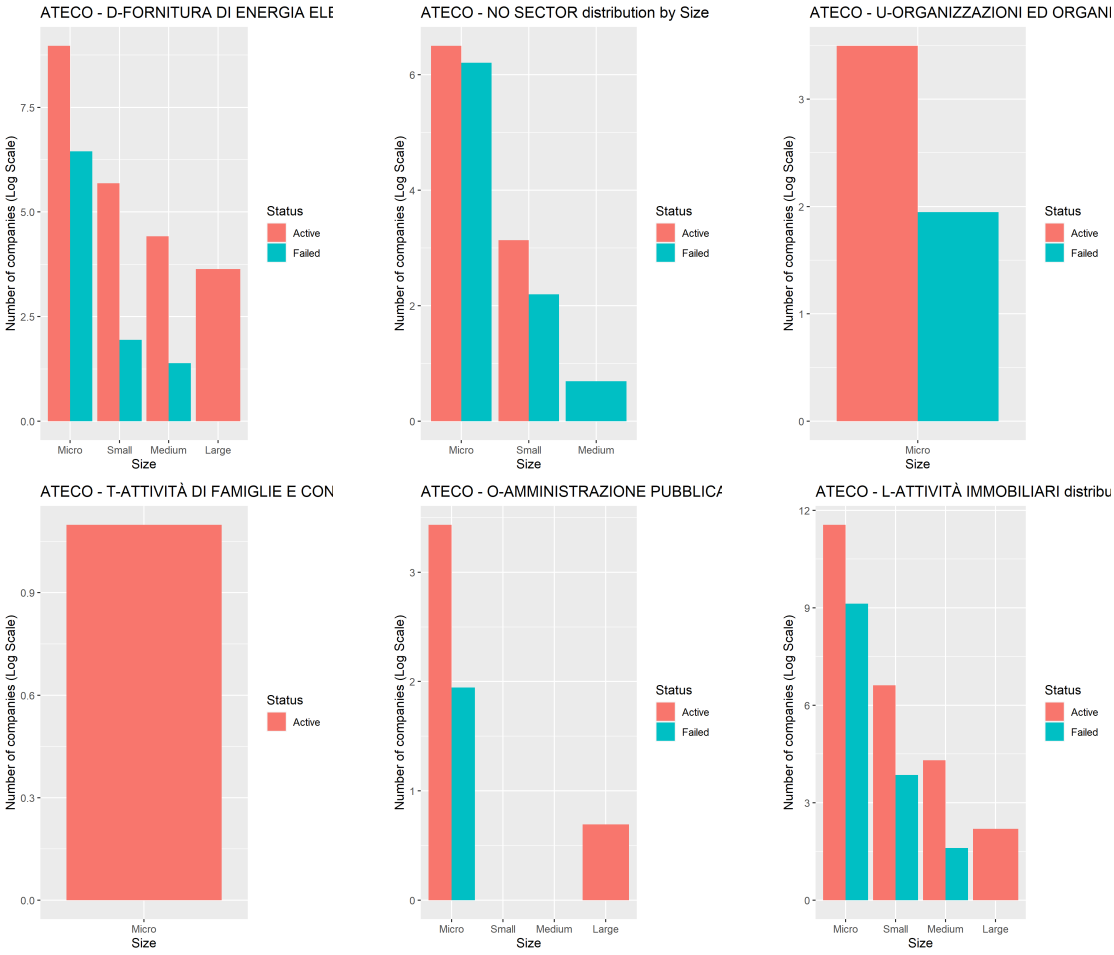


Figure 5: Active failed Densities by Size and ATECO

In 5 we saw how the tendency presented on the earlier graphs continues as micro companies were the most present. In the last two lower graphs of figure 5 we can see how they only present large active. For the O sector there are very few records for small or medium but some records for large which was rare for all the distributions we oversaw. The "No Sector" segment was one of the highlights since it presented one of the very few cases where a size for where there were more failed companies than active. It also lacked 2 full sizes completely.

2.2.3 Region

We vizualized the distributions for some of the regions on Figure 6. We could view how the distributions in the most part all presented some records for each division. Mostly we could see how the majority was for active companies which made sense. However we can see in certain regions like Calabria, Abruzzo and Valle D'Aosta there were sections with just active companies. They all presented a similar pattern.

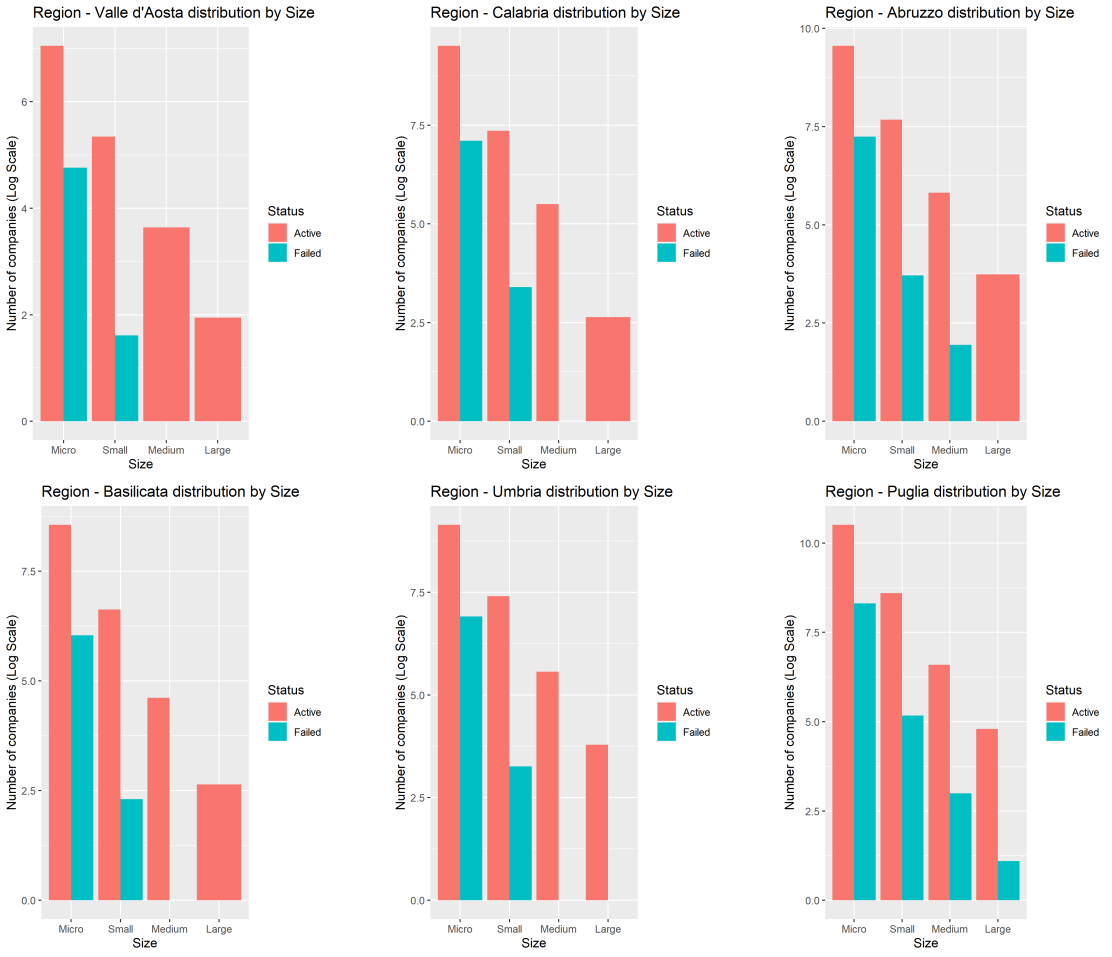


Figure 6: Active failed Densities by Size and Region

3 Comparison for the distributions of size/age/other of failed companies over different years?

This section explains the characteristics of the distributions of the set of companies categorized as failed with respect to different characteristics of time and form with the intention of obtaining a global vision that exposes findings about how long it normally takes for companies to fail or what types of Organizations tend to fail at certain times or under certain conditions. Some relations

that were made to explain these data are, for example, the age distribution of failed companies according to their legal constitution or their commercial sector or the comparison of companies according to the size of their employment and the Italian region where they operated. We opted to analyse the companies report from the years 2015 to 2019 that constitute the majority of data and the most recent cases.

3.1 Analysis by Age

3.1.1 Legal Form

To analyze the relationship of the companies with respect to their age and legal form, the density distributions of the last 5 years were plotted. In general terms, it can be identified that for all sectors it is normal for most companies to fail near the starting years of existence, producing distributions skewed to the right, as can be seen in the figures in the Annexes section. This follows with what we learned from the first questions.

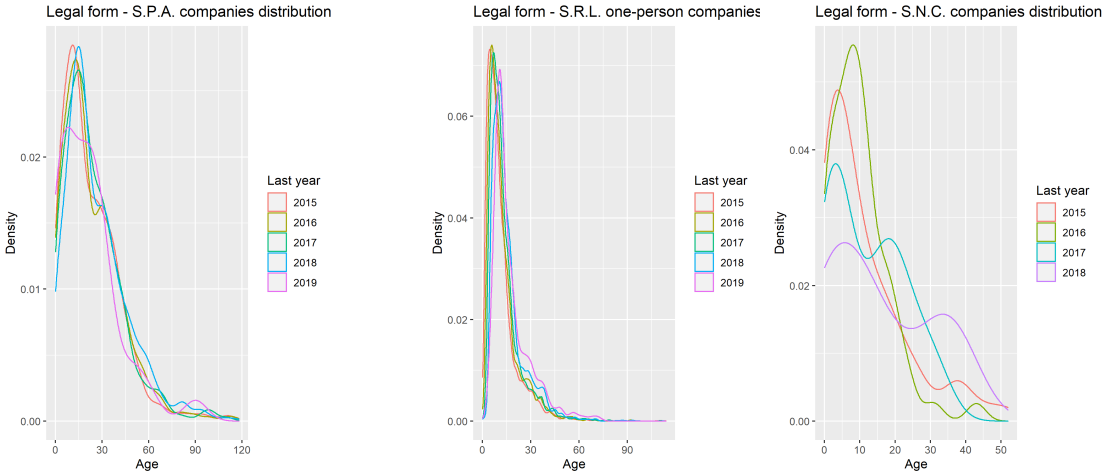


Figure 7: Failed companies distribution by Legal Form and Age

Figure 7 shows three of the most interesting distributions. As can be seen, in the first place, companies of the "S.P.A" type have a fairly skewed distribution that accumulates the vast majority of failed companies between 10 and 20 years of existence and with an abrupt decrease from ages 15 to 60. This behavior is consistent in the 5 years of data collected, so it can be indicated that the second decade of this type of company is crucial for its survival.

Continuing in figure 7, the case of sole proprietor S.R.L companies is quite similar to the previous one, with a positive skew distribution, clearly indicating how practically all companies of this type perish before 15 years of operation. Finally, the companies "S.N.C." They present a somewhat scattered distribution, with apparent bimodal characteristics in some years. However, this bimodal formation varies from year to year, so it is possible that there are errors in the data or that there is not enough data.

Legal Form / Years	2015	2016	2017	2018	2019
Consortium	0.073304	0.027578	0.096973	0.096973	0.16808
S.A.S.	0.16561	0.1363	0.20362	0.20362	0
S.C.A.R.L.	0.040586	0.10522	0.064675	0.064675	0.42766
S.C.A.R.L.P.A.	0.025181	0.065874	0.10354	0.10354	0.18803
S.N.C.	0.057947	0.039069	0.05917	0.05917	0.21669
S.P.A.	0.057947	0.039069	0.05917	0.05917	0.21669
S.R.L.one-person	0.081067	0.084555	0.13075	0.13075	0.13788
S.R.L.simplified	0.10961	0.10937	0.10391	0.10391	0.099847
S.R.L.	0.027979	0.024659	0.11391	0.11391	0.052223
Social cooperative company	0.13182	0.14706	0.068013	0.068013	0.36061
Other	0.24302	0.16885	0.26491	0.26491	0

Table 4: Results for KS test for Question B according to age by Legal Form

The records indicating 0 are records that didn't have enough records to be measured.

3.1.2 ATECO

This section analyzes the distributions of failed companies in relation to the ATECO classification to which they belong. For this, the densities of failed companies in the last 5 years have been plotted for each code of the classification.

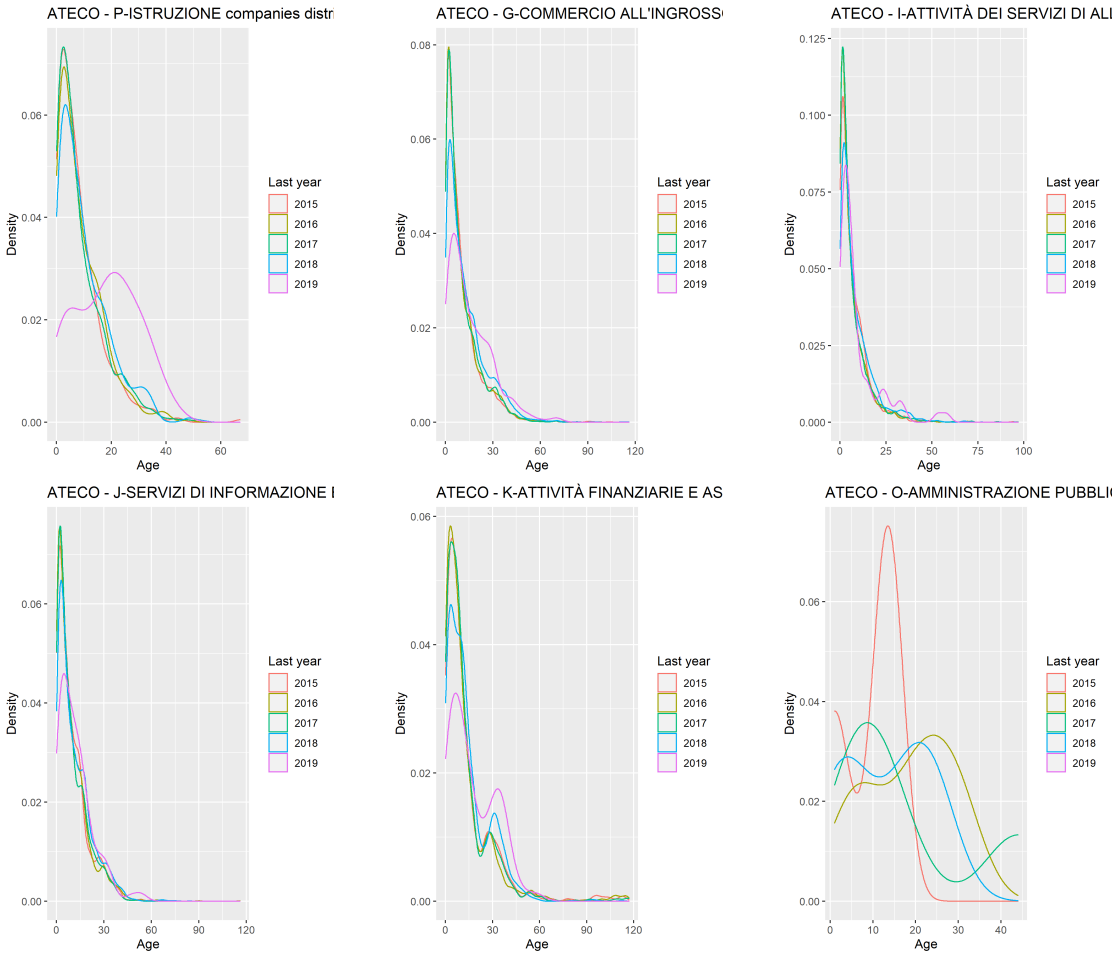


Figure 8: Failed companies densities by ATECO per year

Most of the companies, regardless of the sector or the year in which they are evaluated, present similar characteristics in their density distribution, with a certain tendency towards a positive bias, indicating failures at an early age. Some example cases are shown in figure 8, as seen for companies in sector G and I, respectively, with a high density of failed companies before 15 years of existence. However, some other of these density distributions suggest some interesting behavior. Such is the case of companies in the P sector, as can be seen in the figure, the failure trend of these companies during the 2015-2018 period is practically identical, with a life expectancy clearly less than 10 years. However, for the year 2019, an anomaly appeared that broke the trend, generating a bimodal distribution with a greater number of failed companies, not only in the life range of 0 to 10 years, but also around 20 years. In a similar sense, it can be seen that for public administration companies (sector O), which despite being a rather convulsive distribution with great separation between years (see table 5), there was a peak in density during 2015, which could suggest that some important public measure has taken place. led to the closure of several state-owned companies.

Sector / Years	2015	2016	2017	2018	2019
A	0.037413	0.042117	0.12708	0.12708	0.26195
B	0.11935	0.13607	0.20631	0.20631	0.30862
C	0.018662	0.013781	0.10695	0.10695	0.13645
D	0.098574	0.18587	0.17468	0.17468	0.62372
E	0.09445	0.13575	0.10838	0.10838	0.35105
F	0.04186	0.079292	0.12234	0.12234	0.073373
G	0.014758	0.016042	0.10222	0.10222	0.096515
H	0.021418	0.032449	0.12267	0.12267	0.18016
I	0.055042	0.025494	0.1056	0.1056	0.072654
J	0.032025	0.028354	0.085741	0.085741	0.07687
K	0.03865	0.036238	0.11526	0.11526	0.13213
L	0.058066	0.03123	0.10968	0.10968	0.14801
M	0.0316	0.019399	0.092463	0.092463	0.21735
N	0.032104	0.028659	0.078311	0.078311	0.093287
O	0.66667	0.41667	0.42857	0.42857	0.85714
P	0.049705	0.037707	0.097563	0.097563	0.38782
Q	0.054458	0.032126	0.075195	0.075195	0.14818
R	0.023389	0.030757	0.10887	0.10887	0.15614
S	0.033237	0.029699	0.037928	0.037928	0.18522
NO SECTOR	0.11372	0.36526	0.26555	0.26555	0.31502

Table 5: Results for KS test for Question B according to age by Sector

3.1.3 Region

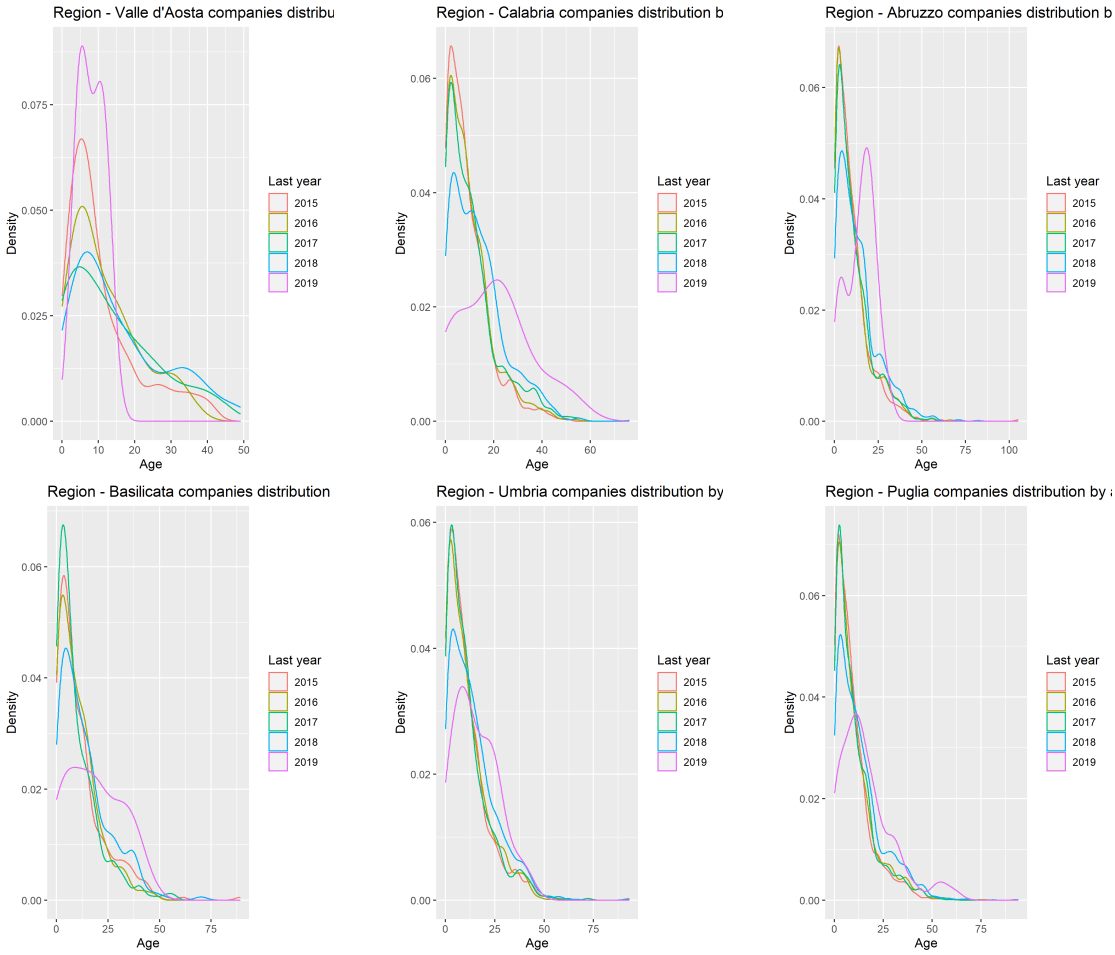


Figure 9: Failed Companies Densities by Region

In the case of the distributions of the failed companies according to their age and depending on their location among the different regions of Italy, we can see in figure 9 that in general terms the density lines also follow the inverse Gaussian figure apparently biased towards the right. Across regions, many distributions are very similar, with the same trend of companies failing in the first few years of operation. The KS tests also do not indicate worthwhile changes in the distributions over time.

Region / Years	2015	2016	2017	2018	2019
Abruzzo	0.023819	0.038835	0.11751	0.29339	0.29339
Basilicata	0.048006	0.085652	0.18843	0.30448	0.30448
Calabria	0.03757	0.048714	0.13787	0.36127	0.36127
Campania	0.02268	0.043455	0.12606	0.12665	0.12665
Emilia-Romagna	0.02511	0.023678	0.09416	0.10942	0.10942
Friuli	0.052766	0.025757	0.13121	0.29232	0.29232
Lazio	0.030093	0.020459	0.1086	0.078649	0.078649
Liguria	0.042375	0.024497	0.10368	0.12251	0.12251
Lombardia	0.024169	0.017136	0.10292	0.075125	0.075125
Marche	0.036969	0.048168	0.13059	0.2059	0.2059
Molise	0.054984	0.074332	0.13566	0.67308	0.67308
Piemonte	0.028907	0.020788	0.098559	0.16967	0.16967
Puglia	0.042942	0.015805	0.13066	0.16698	0.16698
Sardegna	0.029135	0.049729	0.13129	0.28155	0.28155
Sicilia	0.022999	0.036798	0.12804	0.15141	0.15141
Toscana	0.010491	0.036132	0.096574	0.095777	0.095777
Trentino	0.027773	0.049715	0.081609	0.40646	0.40646
Umbria	0.033057	0.041462	0.14274	0.13909	0.13909
Valle	0.12127	0.11165	0.13934	0.47541	0.47541
Veneto	0.042342	0.023047	0.096741	0.10154	0.10154

Table 6: Results for KS test for Question B according to age by Region

3.2 Analysis by Size

3.2.1 Legal Form

Regarding the distributions of failed companies with respect to their classification of size and legal form in the last 5 years of collected data, what is most striking is that the density of failed micro companies is considerably higher than the other categories of failed companies. size, and this is a constant for all legal forms evaluated as shown in figure 10. Another characteristic that could be identified is that for the year 2018 the density of failed micro enterprises is much higher compared to the other years, and is constant for most legal forms, except "S.N.C" and "Other". In the same sense, for the years 2016 and 2017 the distance between distributions of both types of "S.R.L", and others such as "S.P.A" and "Consortium" is low.

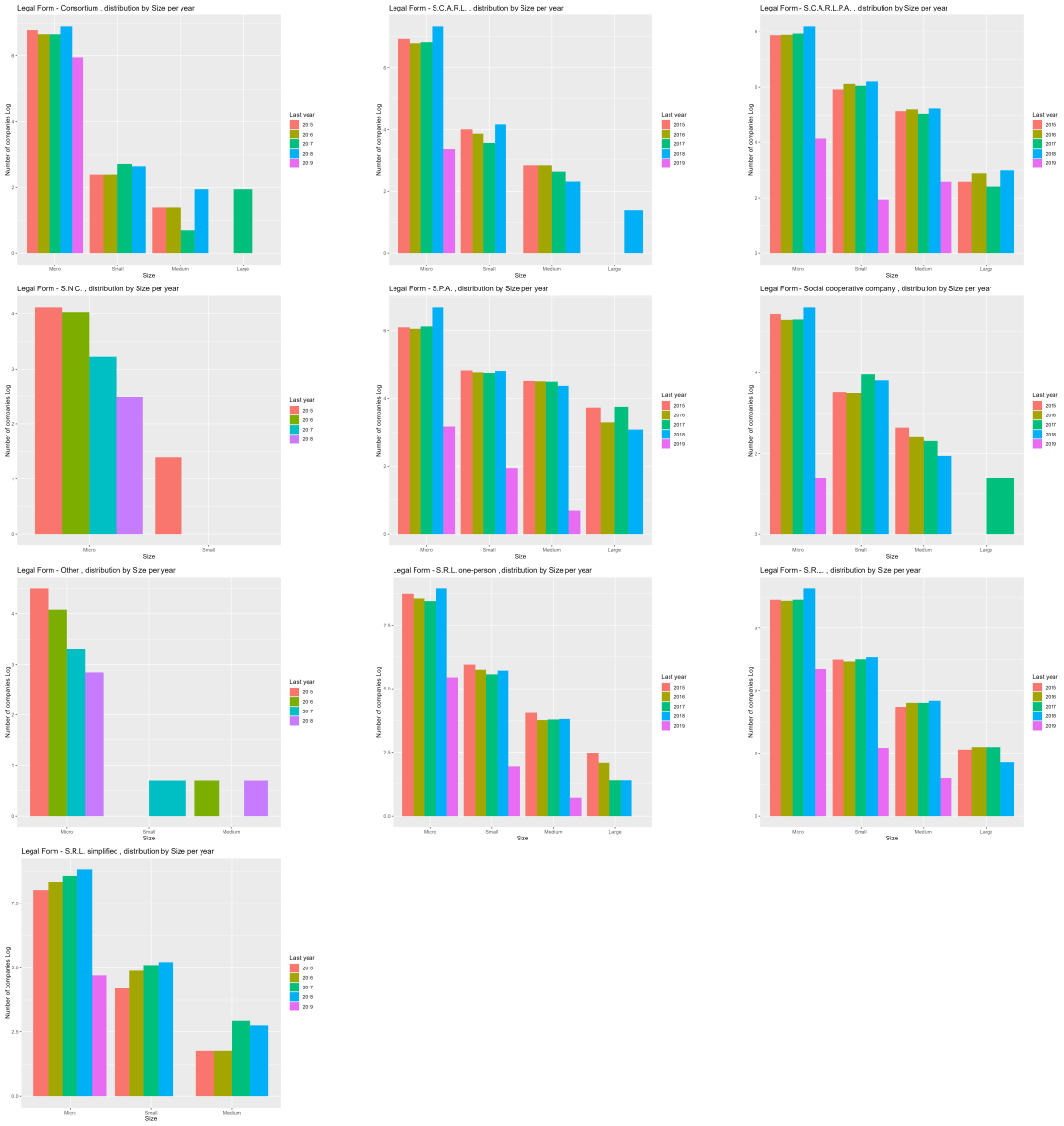


Figure 10: Failed companies by Size and by Legal Form over time

3.2.2 ATECO

In the case of the distributions of failed companies with respect to the classification by ATECO sectors and size in the last five years of data collected, we can see in figure 10, the distributions between the different sectors present very similar characteristics, almost identical in the composition of density by size and even year by year with an apparent reverse trend and as expected given the data seen in the previous sections, a preponderance in the density of failed micro-enterprises in almost every year.

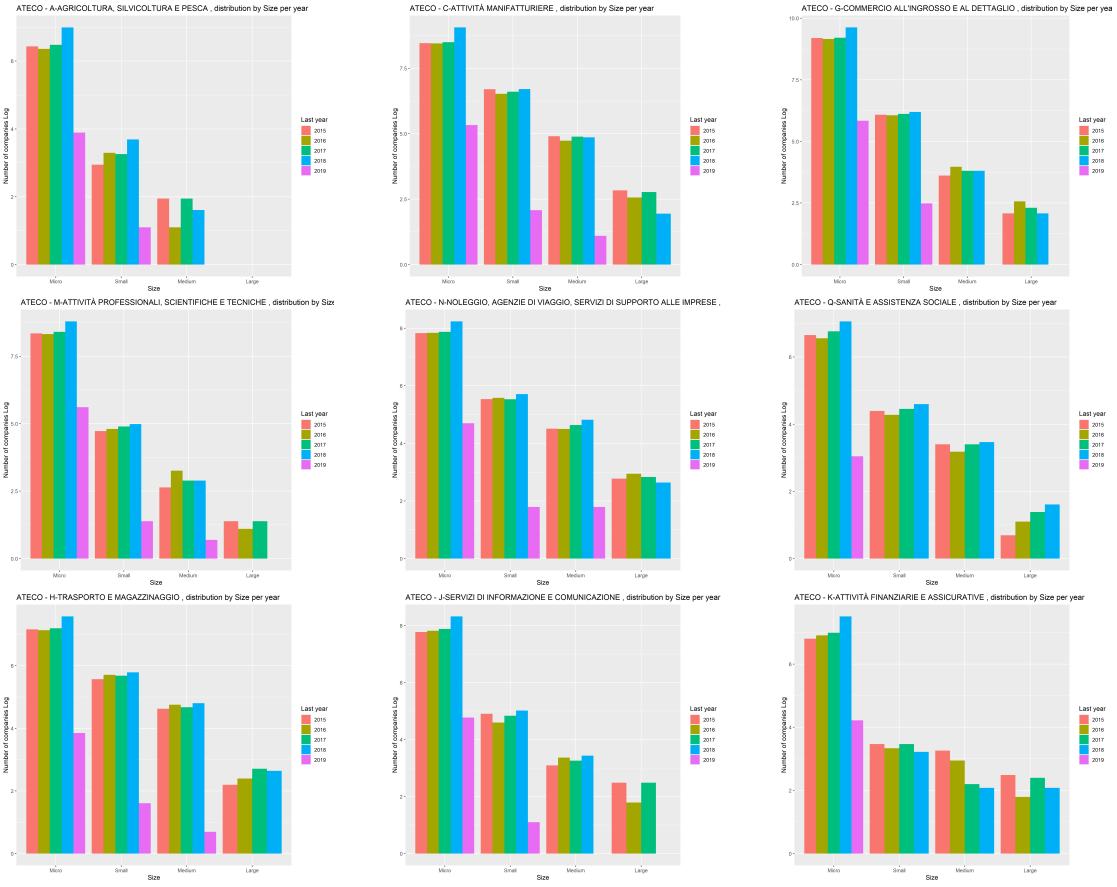


Figure 11: Failed companies by Size and ATECO over last 5 years

3.2.3 Region

To evaluate the distributions of failed companies with respect to their sizes and geographic locations over the years, we decided to group the regions according to their cardinal location on the map, associating each region to one of these four groups: north, center , south and the islands. This was done to try to understand in a less granular way the characteristics of these distributions, although we have included the distributions for each individual region in the annexes.

Analyzing the distributions by their location, no really interesting patterns are identified. In figure 11 we can see similarities between the groups regardless of the geographic region, maintaining the inverse trend for the different sizes of companies and for the different years between which the data is distributed, with 2018 clearly being the year with the highest density of failed companies. regardless of location or size.

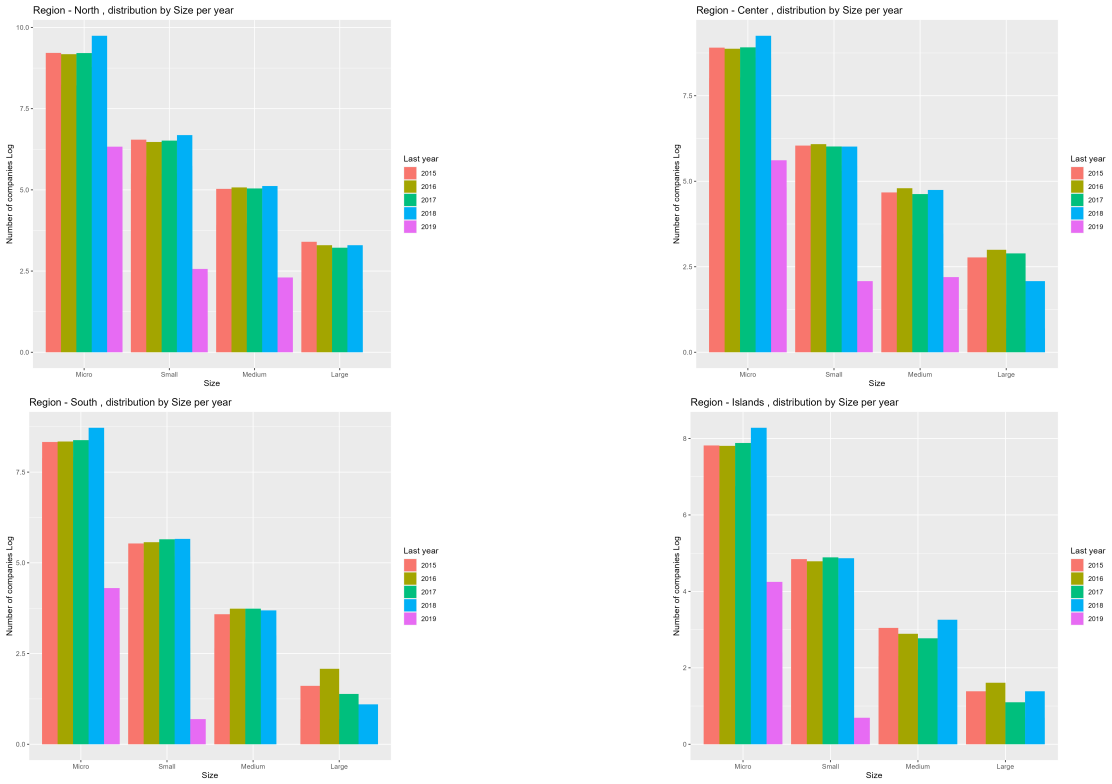
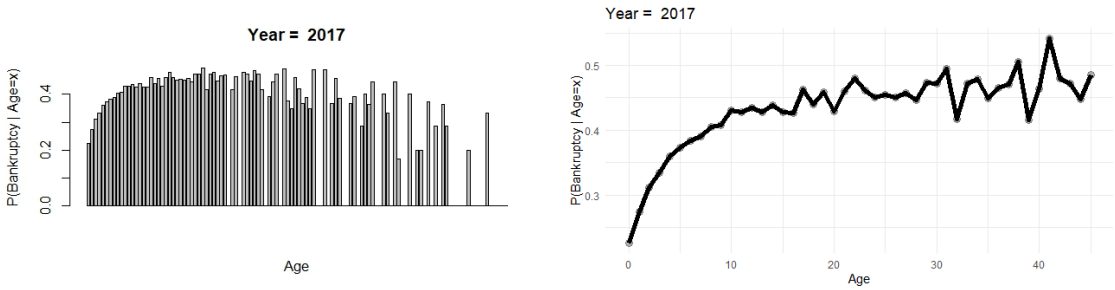


Figure 12: Failed companies by Regions Location and Size over the last 5 years

4 Probability of failures conditional to size/age/other of firms at a specific year

To answer the third question required analyzing the probability of bankruptcy at a specific year, conditional to the age or size of the companies. For the analysis we have chosen the year 2017, we can observe how the probability of failure changes with respect to size, age and sector in which the company operates. When the analysis is deepened by distinguishing the sectors, regimens and company forms, the calculation of the conditional probability must also take this information in consideration, the division of bankrupt companies by the total quantity of the companies successful and failed given the selected year.



(a) (b)

Figure 13: 2017 analysis

The year 2017 it contains a good amount of data for all ages. From Figure 13 a) we can tell the range of ages of the companies that goes from 0 to 117, we can observe that the conditional probability changes with pass of time, because we have fewer companies that succeed while they get through the years, we can tell that while the company gets older, the probability of going bankrupt goes lower, we could say that with increasing age the probability of failure tends to decrease on average. We have some peaks but that is because for companies with ages greater than 100 years are few, for example we have in total 3 companies with 117 years, two that are still active and one that went bankrupt.

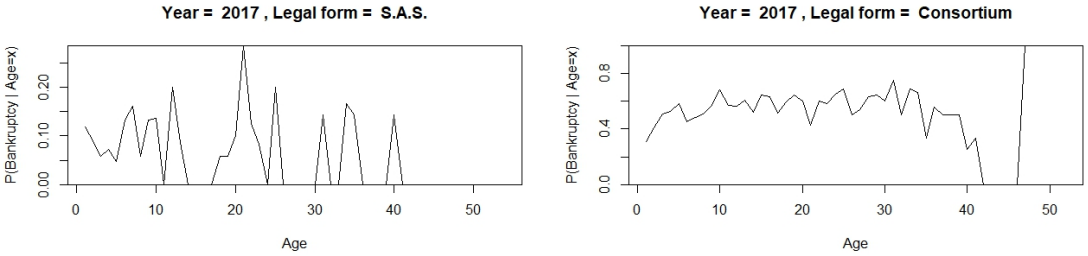
Company_age	Active	Failed	Total	Prob_Failed
117	2	1	3	0.3333333
116	2	2	4	0.5000000
115	1	1	2	0.5000000
114	1	1	2	0.5000000
111	0	4	4	1.0000000
110	3	3	6	0.5000000
109	4	1	5	0.2000000
108	0	2	2	1.0000000
107	2	2	4	0.5000000
106	2	2	4	0.5000000
105	0	1	1	1.0000000
102	1	1	2	0.5000000
101	0	1	1	1.0000000
100	1	1	2	0.5000000
99	0	1	1	1.0000000
98	1	2	3	0.6666667
97	10	4	14	0.2857143
96	7	4	11	0.3636364
95	1	1	2	0.5000000

Figure 14: Oldest Companies

4.1 Legal form

4.1.1 Age

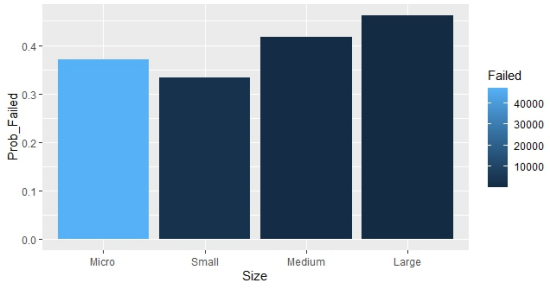
In this section we took in analysis the company form, in order to see if we find that the conditional probability changes or not with respect to this factor, there is a prevalence like we have see in the general part of the question , we can see that there is a more quantity of younger companies that are active with fewer bankruptcy cases, we can notice the same effect of time, while the time passes there are fewer companies that are still active. For example if we emphasize on the company form Consortium and S.A.S we can see in the graph how it begin to decrease the amount of companies that are successful more or less after the 40 years of age From 60 a long time up to 75 companies tend to not come up short, whereas after 75 there's a unused crest within the thickness with likelihood up to 100% of failing. Regarding "Others" (it incorporates SAS and SNC), after a to begin with top within the to begin with 10/15 a long time, the trend decreases altogether. We have two unused crests over 0.50 at



(a) (b)

Figure 15: 2017 S.A.S and Consortium probability graphs

4.1.2 Size



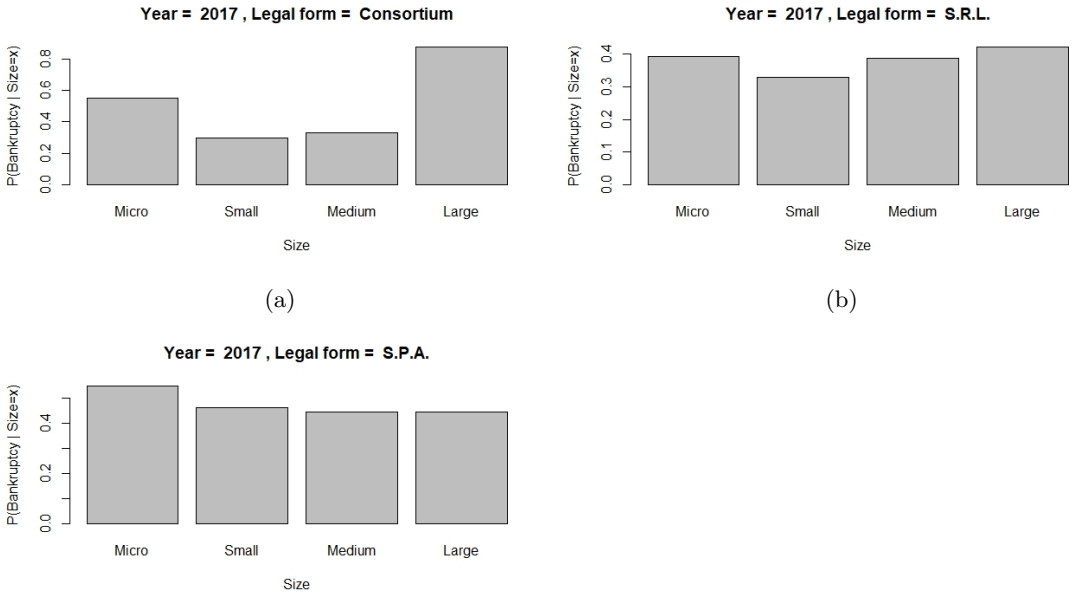
Size	Active	Failed	Total	Prob_Failed
Micro	80179	47315	127494	0.3711155
Small	5860	2932	8792	0.3334850
Medium	789	564	1353	0.4168514
Large	113	97	210	0.4619048

(a) (b)

Figure 16: Probability of failure by size

With respect to the Size, we have to remember that the is a higher quantity of companies with an Active status than Failed, from the general graph of the size of the companies we can see that for the year 2017 we have the same repercussion the Larger companies have a greater probability to fail than the smaller companies. Focusing on the legal form aspect of the companies we can

observed different aspects, we can see that the conditional probability of going bankrupt changes, for example: The legal form S.R.L and Consortium have a higger probability of failure for the larger companies, that is the opposite case for the legal form S.P.A



(a) (b)

(c)

Figure 17: Probability of failure on 2017 by size

4.2 Industry sector (ATECO)

4.2.1 Size

Analyzing the distributions of the probability of failure of companies according to their size and their economic sector, it was possible to see that each group is not necessarily homogeneous with the other. Some interesting comparisons can be seen in figure 18, which shows how companies in the hotel and restaurant sector, as well as financial and insurance companies, have a greater probability of failure when they are large companies; while professional and scientific activity companies, or construction companies have a greater probability of failure when they are smaller.

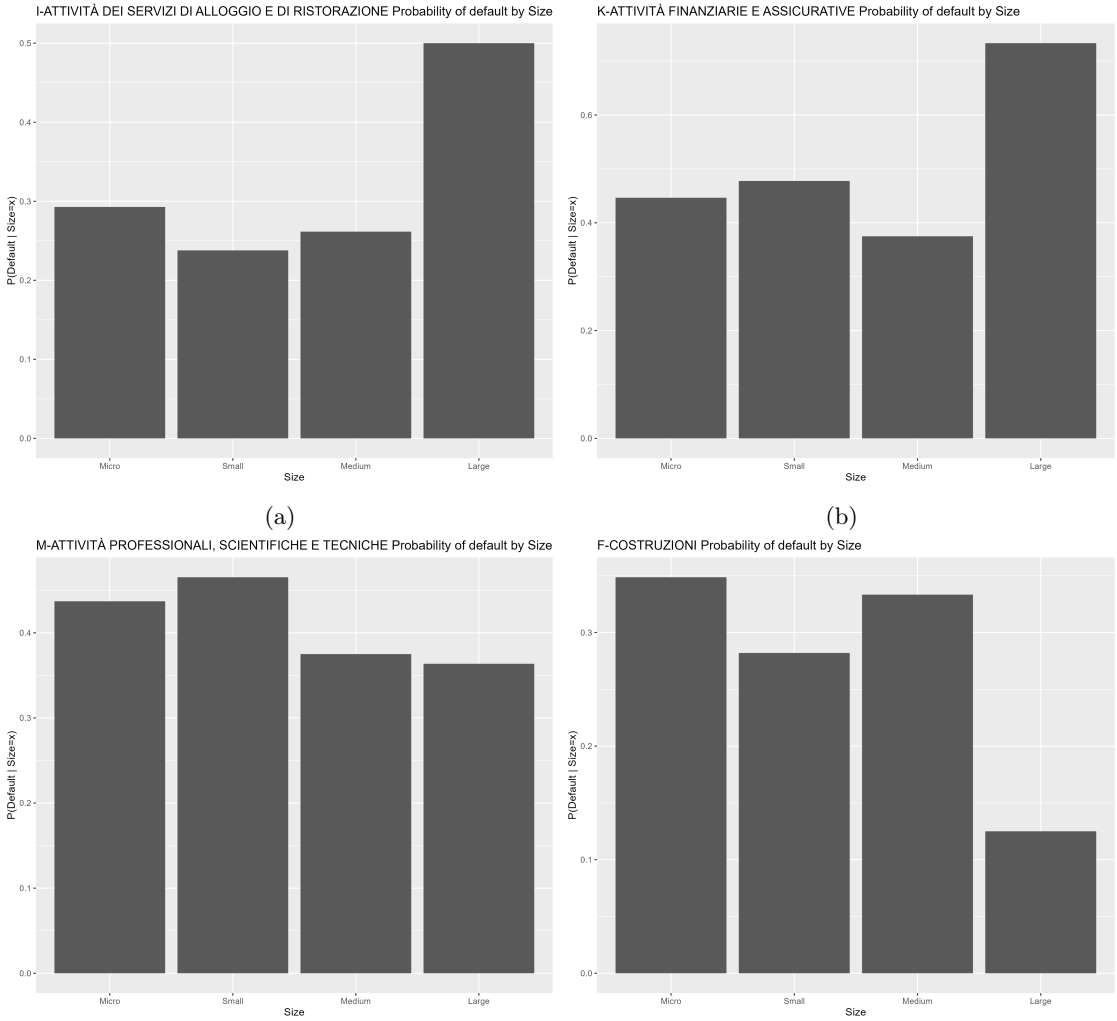


Figure 18: Probability of failure on 2017 by ATECO and by size

4.2.2 Age

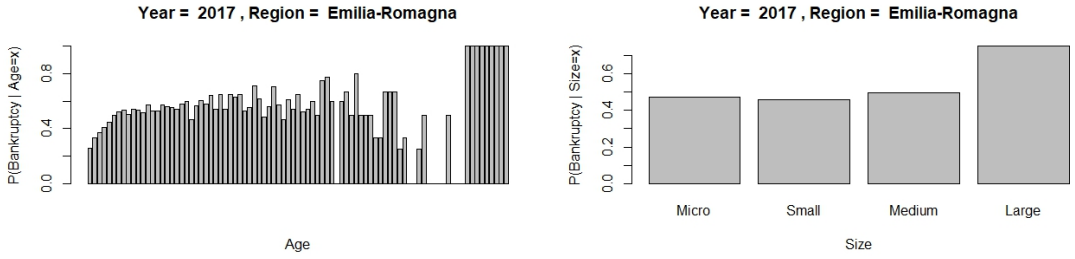
Analyzing the companies with respect to the age and the sector where they belong. As it is expected, the distributions tends to have peaks in probability when companies are young. Since the sample for young companies is greater, the probability of dafault gets more irregular (noisy) with the years. Also, diverse ATECO sector doesn't have many failed old companies, which can help to explain the tail of the distribution. In figure

(a) (b)

Figure 19: Probability of failure on 2017 by ATECO and by age

4.2.3 Region

As shown in this entire section, similar behavior is found in the distribution of probabilities with respect to the ATECO and the location of the companies. An example is region Emilia-Romagna, as it can be seen in figure 20 companies risk of failure increase with age, but after a certain time, the probability of going default begins to decrease, from the barplot of the size companies for the region Emilia-Romagna can see that Large companies have greater risk to fail, but this can be related to the fact that there are few large companies. Other distributions can be see in Annexes.



(a) (b)

Figure 20: Probability of failure on 2017 by ATECO and by Region

5 Fitting models for failure prediction

This section aims to predict whether a company will end up failing using parametric and non-parametric algorithms. For this, a binary scoring model or classification system will be modeled, based on the probability of failure given the available variables. Later on, a rating model or multiple ordinal class model is also developed to predict probability of certain companies solvency health classification.

5.1 Data Preparation

In order to create the models, it was necessary to apply some additional procedures to the dataset worked on in section 1. First, we omit all null values, subtracting only the records that are complete. This left us with a dataset of 214,238, made up of 21 continuous variables from all the financial indicators and the class variable.

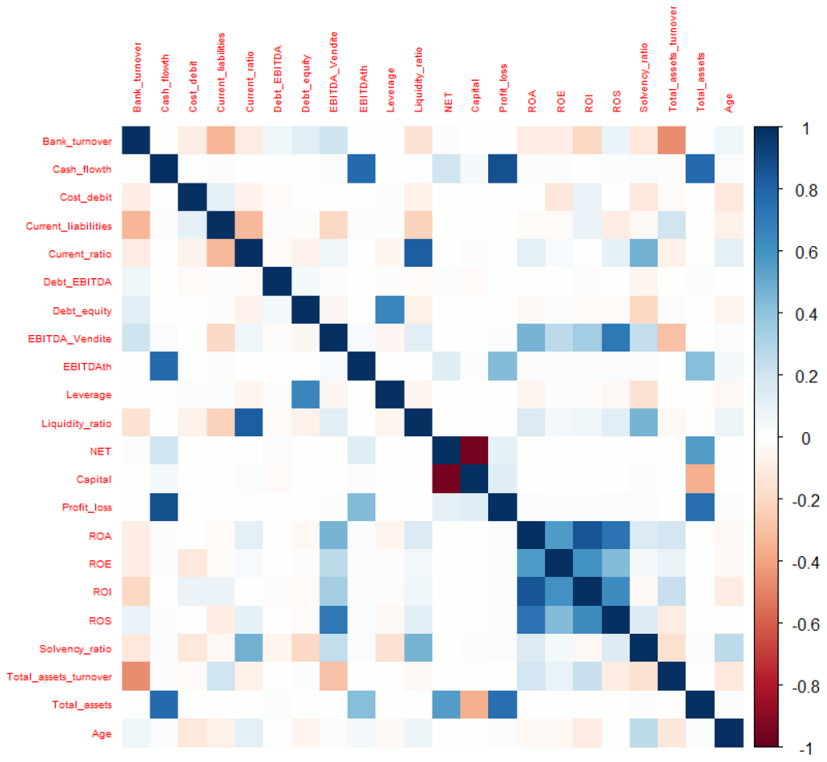


Figure 21: Features correlation heatmap

5.1.1 Correlation Analysis

Secondly, we performed correlation analysis in order to find redundant features. In this sense, the Pearson correlation index between each feature was calculated, and then plotted in a heatmap to ease the interpretation. As it can be seen figure 21, some variables have characteristics of high correlation. For this reason, we decided to eliminate the variables with a correlation equal to or greater than 0.8 according to the Pearson correlation index. In this sense, the following variables were removed:

- ROA
- Cash Flowth
- NET
- Liquidity Ratio

5.1.2 Data splitting and Balancing

To evaluate the performance of a machine learning algorithm, it is needed to split the data. A subset of the dataset must be considered for training so the algorithm learns how to predict future instances, and then another subset is used to test the model, to examine how good the algorithm is predicting unseen samples. Usually, taking as the test datatset 20% of the dataset

Age Range	Active	Failed	Total	Prob_Failed
30 or more	33887	3022	36909	0.08187705
less than 30	138711	38618	177329	0.21777600

Figure 22: Caption

is a good approach, however as we would like to consider old companies to predict newer ones, it is necessary to perform the split in a controlled way.

After diving into the data distributions, We noticed that 93% of the companies were younger than 40 years, and 80% of them were less than 30. Furthermore, as it was shown in previous sections, the dataset is really unbalanced as the number of active companies is much higher than failed ones, and this behavior is even more serious with the older companies. As it can be seen in figure 22, the relation of active/failed companies in 30+ years old companies is about 90/10. For these reasons, we decided to work only with companies younger than 30 years.

The next step was to split the data temporarily. For this, we took all the companies older than 5 years to include them in the train set, and the rest of the companies were left for the test set. This division was carried out conveniently since it allowed obtaining an approximate ratio of 75/25 in the division of the data.

Having already made the division, even the training set had a large imbalance in the classes due to the number of active companies, which would seriously affect the modeling results. For this reason, it was decided to apply SMOTE to generate synthetic examples of the minority class until a slightly more balanced set was obtained. After several tests, we were left with a training set with a ratio of approximately 60/40.

5.1.3 Normality, Variance Homogeneity, and Multicollinearity

Normality

To determine if the training set complies with the principle of normality, it was decided to evaluate the distribution of each variable with respect to the Shapiro-Wilk test, which calculates the percentage of similarity between the observed distribution and the normal distribution. The null hypothesis of this analysis indicates that the distribution is statistically normal, and if we take as a rule that the null hypothesis is rejected if we obtain a p-value less than 0.05, then we can affirm by looking at the results of figure 23 that the distributions of our variable are not normal.

	statistic	p.value
bank_turnover	0.905566	2.9855e-48
cost_debit	0.9570865	4.859974e-36
current_liabilities	0.8454296	8.060796e-57
current_ratio	0.7121147	1.146203e-68
debt_ebitda	0.2261725	7.076882e-90
debt_equity	0.3671636	2.514323e-85
ebitda_vendite	0.8100238	1.273918e-60
ebitd_ath	0.0757994	5.394209e-94
leverage	0.2244888	6.309745e-90
capital	0.1891743	5.969466e-91
profit_loss	0.08361826	8.532716e-94
roe	0.8078773	7.842002e-61
roi	0.9695713	2.697455e-31
ros	0.8539995	8.615308e-56
solvency_ratio	0.8888226	5.208081e-51
total_assets_turnover	0.8983967	1.771158e-49
total_assets	0.1118027	4.586894e-93

Figure 23: Shapiro-Wilk normality test over training set

To try to remedy this a bit, we try to apply some simple transformations on these distributions, such as logarithm or square root. However, the distributions still did not approach normality.

Multicollinearity

This is a phenomenon that occurs when 2 or more variables are very correlated between them, such that they do not provide independent information to the regression model. If the correlation is too high, it can cause problems when fitting and interpreting the model. The most common way to determine multicollinearity is by calculating the variance inflation factor (VIF), which measures strength of the correlation between the variables in from a regression model.

Features	VIF
bank_turnover	1.656269
current_liabilities	1.413299
debt_ebitda	1.008266
ebitda_vendite	2.542347
leverage	2.079934
profit_loss	5.571251
roi	2.768422
solvency_ratio	1.66766
total_assets	7.076867
cost_debit	1.103802
current_ratio	1.676951
debt_equity	2.131865
ebitd_ath	1.225818
capital	2.626479
roe	1.711436
ros	3.295941
total_assets_turnover	1.536173

Table 7: Variance Inflation Factor

In order to have this calculation we first have fitted a regression with all the financial indicators of the data-set without regularization. Then we calculated the correlation between this attributes and then we proceed to obtain the variance inflation factor. The result of this analysis can be seen in the table 7, where we can identify the variables "Profit" and "Total Assets" with a VIF value above 5, indicating a possible severe correlation with another variable.

Variance Homogeneity

To determine the homogeneity of the variance, the most common is to use the F-test, however this assumes that there is normality in the distributions. For this reason, we decided to use the Fligner-Killeen test, which is a non-parametric test but is much more robust against distributions far from normality. Even so, when applying this test we verify with very low p-values that the homogeneity of variance is not fulfilled for these distributions.

5.2 Scoring Models

For the scoring models, it was decided to use 3 methods, Linear Discriminant Analysis (LDA), Boosted Logistic Regression (LR) and Random Forest (RF). The first two were selected because they are parametric algorithms widely used for prediction tasks. For its part, Random Forest is a non-parametric algorithm that works quite well with any type of classification task, so it is useful for comparison purposes.

The training of the algorithms began using the final data training set, having also included the scaling and centralization of the data to adjust LDA and LR as parametric algorithms, although these algorithms assume the presence of characteristics such as normality, variance homogeneity and the multicollinearity of the distributions and in the previous section it was proved that our data does not meet these characteristics, so we only adjust it for academic purposes. In the case of the LDA, as we see in table 8, the linear discrimination coefficients and the mean groups that make up the linear combination of the input variables for model decision making were extracted.

Of these data, no variable has a considerable influence on the final classification, but all of them contribute to some extent to the maximization of the separation between the groups.

LDA Group Means		
Variables	**Active**	**Failed**
bank_turnover	-0.11166	0.164309
cost_debit	-0.20727	0.305003
current_liabilities	-0.13394	0.1971
current_ratio	0.073295	-0.10786
debt_ebitda	-0.00321	0.004717
debt_equity	-0.0714	0.105062
ebitda_vendite	0.077539	-0.1141
ebitd_ath	0.00983813	-0.01447739
leverage	-0.05384	0.079232
capital	0.000246	-0.00036
profit_loss	0.007747	-0.0114
total_assets_turnover	0.074412	-0.1095
total_assets	0.004125	-0.00607
roe	0.16819	-0.2475
roi	0.113633	-0.16722
ros	0.09539	-0.14037
solvency_ratio	0.121382	-0.17862

Coefficients of linear discriminants	
Variable	**LD1**
bank_turnover	0.462929
cost_debit	0.604099
current_liabilities	0.586677
current_ratio	0.230902
debt_ebitda	-0.00551
debt_equity	0.113108
ebitda_vendite	0.01167
ebitd_ath	0.002604
leverage	-0.00399
capital	0.007232
profit_loss	-0.00913
roe	-0.25072
roi	-0.15189
ros	-0.07949
solvency_ratio	-0.29345
total_assets_turnover	-0.06972
total_assets	0.000877

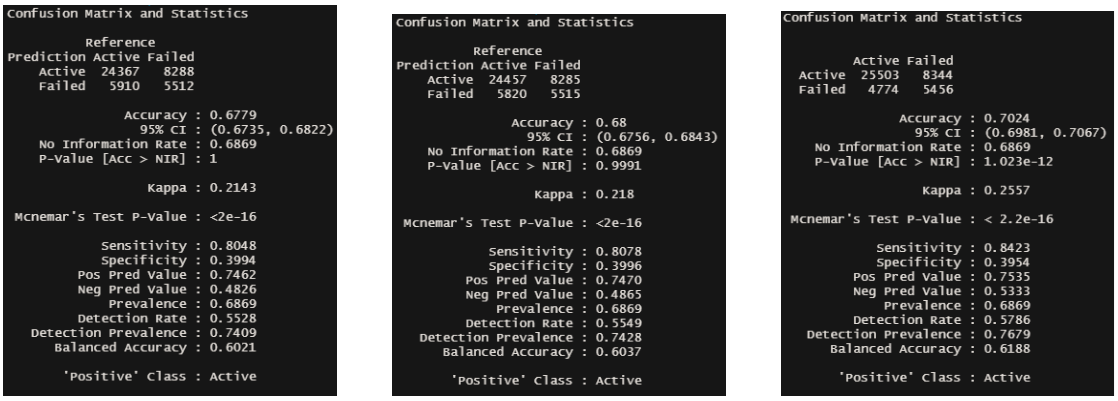
Table 8: LDA Fitting parameters

The second adjusted model was the logistic regression, for which we can find its calculation parameters in table 9. From these data, some important findings are the variables with p-value greater than 0.05 such as "capital, "profit_loss" and the three variables related to ebitda. These high values indicate that they are not statistically significant for the model. Then, seeing coefficients, we can see that the variables with the greatest influence on the model are current_liabilities and cost_debit, and since they are negative, this means that their increase is associated with approximately a 50% reduction in the probability that the company will fail. Seeing the positive coefficient of the intercept, we can know that the probability of predicting that the company will not fail is above 50%.

To evaluate the performance of the models shown above, we have used cross-validation and metrics such as accuracy and Kappa. In figure 24, it can be seen the results of the prediction of each of the models, and in general terms the predictions of the three models were similar and overall quite bad. Appreciating the very high sensitivity and low specificity, this means that the models have high recall but very low true negative rate. From the data we have, it can be concluded that the models are only correct about 40% when predicting that the company would fail. Also, looking at the confusion matrix shows the high number of false positives for each of the models.

	Estimate	Std. Error	z value	Pr(>—z—)
(Intercept)	0.449021	0.005219	86.03	<2e-16
bank_turnover	-0.39144	0.00691	-56.651	<2e-16
cost_debit	-0.50713	0.0055	-92.203	<2e-16
current_liabilities	-0.52832	0.006677	-79.126	<2e-16
current_ratio	-0.20578	0.006981	-29.477	<2e-16
debt_ebitda	0.005786	0.005119	1.13	0.258
debt_equity	-0.17598	0.011985	-14.684	<2e-16
ebitda_vendite	-0.00335	0.008427	-0.398	0.691
ebitd_ath	-0.00945	0.009346	-1.011	0.312
leverage	0.019969	0.008848	2.257	0.024
capital	-0.05204	0.06741	-0.772	0.44
profit_loss	0.024174	0.024623	0.982	0.326
roe	0.204749	0.00702	29.166	<2e-16
roi	0.134166	0.008742	15.347	<2e-16
ros	0.072399	0.009646	7.505	6.12E-14
solvency_ratio	0.254556	0.007083	35.94	<2e-16
total_assets_turnover	0.063038	0.006527	9.658	<2e-16
total_assets	0.011237	0.029072	0.387	0.699

Table 9: Logistic Regression Parameters



(a) LDA (b) Log Regression (c) Random Forest

Figure 24: Scoring models prediction results.

To remedy this, there are many techniques that can be applied, and one of them would be to modify some parameter of the algorithms so that they can better capture the class that we want to prioritize. For this we must make the specificity increase, increasing the TNR and therefore reducing false positives. By plotting the receiver operating characteristic curve (ROC) for each of the models, as seen in figure 25, we can see the relationship between sensitivity and false positive rate, indicating the coverage of the model.

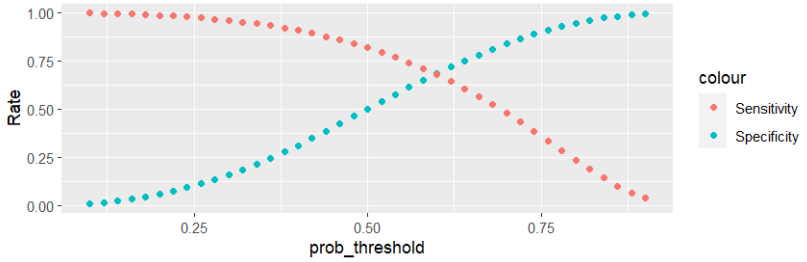


Figure 26: Logistic Regression Sensitivity and Specificity given a probability threshold

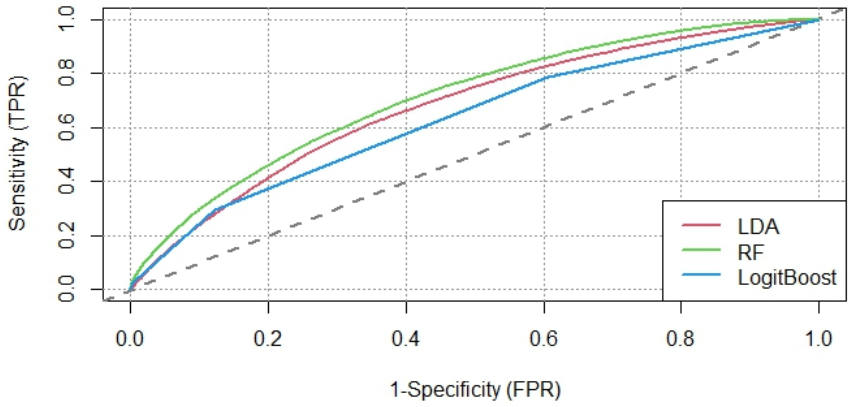


Figure 25: Models ROC curve

If it is taken the logistic regression model as an example, we know that it has a threshold that serves as a parameter for the final decision made by the model and that in principle it is configured as 0.5. Then, looking at figure 26 that shows how the specificity increases as the model threshold increases, we find a point that would give our model a better result, which is approximately 0.625. we could then modify this parameter to reduce false positives.

]

5.2.1 Uncertainty from Confidence Intervals

As the values of the standard error were presented in table 9, those can be used as a measure of the uncertainty of the coefficients of the logistic regression. In this sense, it was decided to calculate the confidence intervals for each of the coefficients of the model at 95%. For this, the following formula was used: $\exp(B \pm 2 \times SE)$

Table 10 shows the confidence intervals for the regression model that was adjusted, and from these data some findings arise that can be raised, such as:

- Based on the data, an increase of around 50% in the odds of a company failing can be expected depending on the value of the cost debit.

- There is a 95% confidence that current liabilities increase between 51% and 54% the odds of a company failing.
- There is sufficient confidence to affirm that the solvency ratio of a company decreases the odds of it failing between 11% and 15%. Likewise, the ROI also decreases the odds of failure by 24% and 26%.

	2.50%	**97.50%**
(Intercept)	0.438796	0.459256
bank_turnover	-0.405	-0.37791
cost_debit	-0.51791	-0.49635
current_liabilities	-0.54142	-0.51525
current_ratio	-0.21945	-0.19209
debt_ebitda	-0.00424	0.015839
debt_equity	-0.19962	-0.15262
ebitda_vendite	-0.01986	0.013177
ebitd_ath	-0.03248	0.005705
leverage	0.001901	0.036742
capital	-0.19626	-0.12533
profit_loss	-0.015	0.084916
roe	0.19101	0.218529
roi	0.117036	0.151306
ros	0.053497	0.091311
solvency_ratio	0.240677	0.268442
total_assets_turnover	0.050253	0.07584
total_assets	-0.02906	0.079823

Table 10: Logistic Regression Confidence Intervals

5.3 Rating Model

For the modeling of the rating system, it was decided to convert the Legal Status variable into an ordinal category as seen in table 11, considering the categories "Active", "Active (default of payments)", "Active (receivership)" as a healthy business, the "Bankruptcy" and "In liquidation" categories as injured businesses, and all other categories as dead businesses. As seen in the distribution of companies according to these categories in the same 11 table, the rating model would be the same with the dataset quite unbalanced.

Legal Status	Rating	Total values
Active	Healthy	141081
Active (default of payments)		
Active (receivership)		
Bankruptcy	Injured	19549
In liquidation		
Dissolved	Dead	19350
Dissolved (merger)		
Dissolved (liquidation)		
Dissolved (demerger)		
Dissolved (bankruptcy)		

Table 11: Rating variables distribution

Using the dataset prepared from the previous section and scaled, a multinomial linear regression was modeled. This algorithm penalizes the differences between the class-specific parameter vectors, instead of penalizing the number of explanatory variables. The model provides interpretable parameter estimates like those of traditional regressions.

When applying this algorithm, we obtain not very satisfactory results. This is to be expected given the imbalance in the data. In the 12 table we can see the coefficients of the Injured and Dead classes being compared with respect to the baseline that are the Healthy companies. In general, according to these data, Total_Assets is the variable whose increase positively influences the probability of the company being classified as a dead company against healthy companies.

Coefficients		
(Intercept)	47.96009	-142.587
Bank_turnover	1.754903	1.713835
Cost_debit	1.930142	1.56601
Current_liabilities	3.099821	3.109758
Current_ratio	1.648455	1.819968
Debt_EBITDA	-0.02168	0.472327
Debt_equity	1.348205	1.422623
EBITDA_Vendite	0.553275	0.377944
EBITDAth	12.12561	4.271332
Leverage	2.068376	2.31223
Capital	0.821336	0.870844
Profit_loss	12.02622	3.613768
ROE	0.09981	0.10666
ROI	0.085278	0.081256
ROS	0.203665	0.156518
Solvency_ratio	0.114591	0.088937
Total_assets_turnover	0.081463	0.067891
Total_assets	3.933999	11.38922

Table 12: Multinomial Logistic Regression Coefficients

When using the model to predict, the coefficients are extracted as seen in table 13, for which we can see that it is very difficult to capture the companies classified as dead, with a very low

sensitivity and detection rate with respect to the other classes. This model is being greatly affected by the unbalance causing a high overfit.

	Healthy	**Injured**	**Dead**
Sensitivity	0.6093	0.60476	0.3447
Specificity	0.7459	0.74724	0.84377
Pos Pred Value	0.8969	0.2257	0.20998
Neg Pred Value	0.3448	0.93946	0.91445
Prevalence	0.7839	0.1086	0.10751
Detection Rate	0.4776	0.06568	0.03706
Detection Prevalence	0.5325	0.29098	0.1765
Balanced Accuracy	0.6776	0.676	0.59423

Prediction	**Reference**		
	Healthy	**Injured**	**Dead**
Healthy	17191	828	1149
Injured	6723	2364	1387
Dead	4302	717	1334

Table 13: Rating Model Prediction