UNIVERSITY OF PISA

MASTER'S DEGREE IN
DATA SCIENCE & BUSINESS INFORMATICS

STATISTICAL METHODS FOR DATA SCIENCE

RISK OF BUSINESS FAILURE

ALESSANDRO BONINI[604482]
CARLO ALBERTO CARRUCCIU[533967]
MARCO CIOMPI [537856]
FRANCESCO SALERNO[534622]

July 22, 2021

Contents

1 Introduction

The objective of this study is to investigate the risk factors of business failure of Italian Companies in the AIDA dataset. In particular, we address 3 business questions regarding different aspects affecting the forementioned risk that will be explained in 3 different sections. The first two aim to categorize the distributions of different types of companies according to their status (active or failed), their age and size, while the last section aims to classify the status of companies using a set of pre-processed features. The code to produce the results of the three questions consists of a series of scripts starting with the loading of the aida.RData dataset.

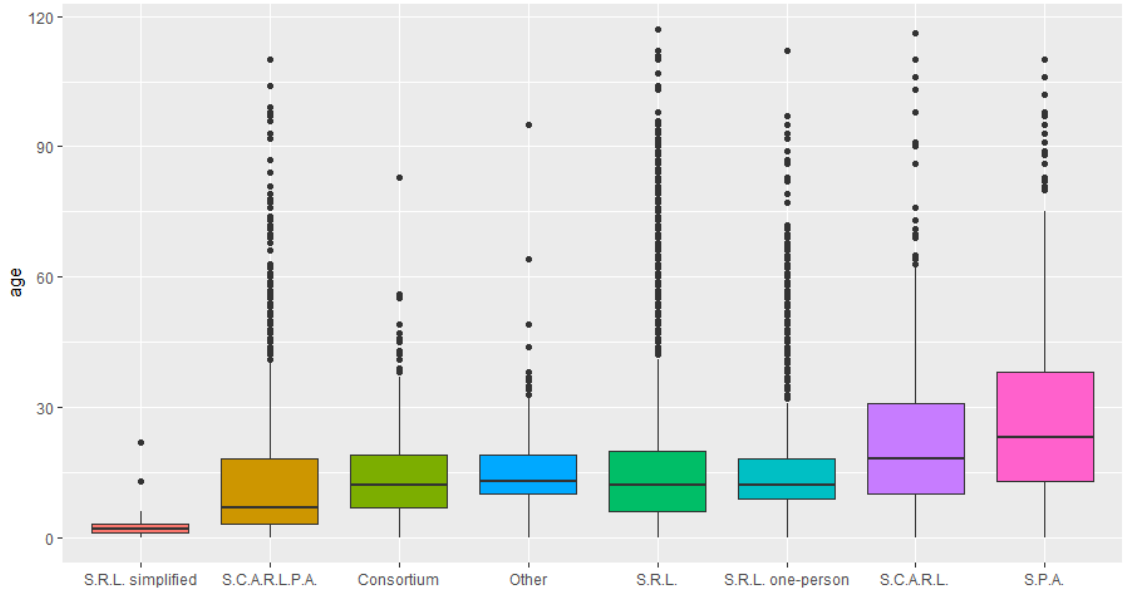


Figure 1.1: Age of Failed Companies by Legal Forms

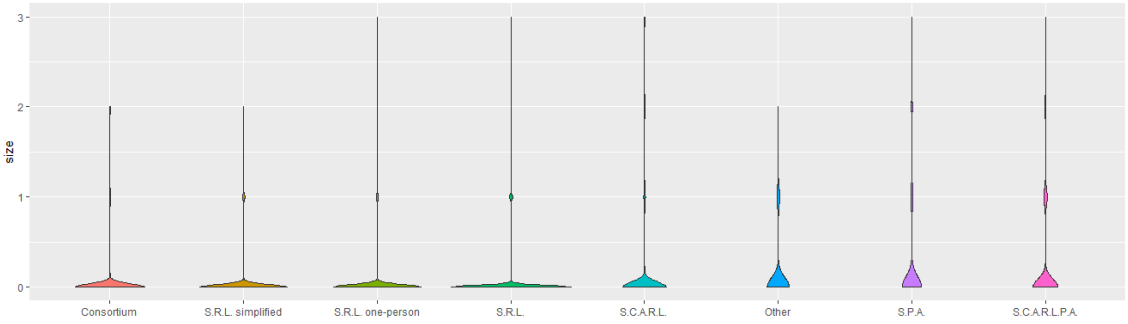


Figure 1.2: Size of Failed Companies by Legal Forms

1.1 Data Exploration

In the part A.1, A.2, was requested to analyse the distribution of ordinal attributes among the target attribute (status of the company); instead, for the point B the distribution of the target attribute over the ordinal ones was analysed. Basically, the analysis were made by graphs and by statistical tests. All codes for the plots and for the tests are available in the relative R scripts. Practically, in the next few sections we're going to describe our analysis of distributions between active and failed companies over aida dataset so as to extract statistical differences among them. In order to answer the questions about the comparison between the distribution of failed and active companies at a specific year or for different years, of on specific features (size/ age), we have started with a cleaning pre-processing of the dataset.

1.2 Data Preparation

Firstly, we started by eliminating null rows from attributes involved to calculate AGE and SIZE (features to answer to question A and B). Moreover, we eliminated the samples having Legal Status equal to either "Dissolved (demerger)" or "Dissolved (merger)" since they relate to companies dissolved in order to be merged in new entities or the contrary; therefore, we cannot know if the company is failed or not. In addition, we reduced the number of unique values in the Legal Form by grouping into Others those values having a relatively small sample.

Then, we focused on creating new features to better analyze the problem at hand. The new features that have been created are the following:

- **Age**, defined as: Last accounting closing date – Incorporation year ;
- **Size**, defined as a categorical variable having 4 possible values and using the attributes Number of employ eesLast avail. Yr and Total assetsth EURLast avail. Yr:
  1. *Micro enterprise* : if Number of employees < 10 and Total assets < 2 mln€
  2. *Small enterprise* : if 10 < Number of employees < 50 and 2 mln €< Total assets < 10mln€
  3. *Medium enterprise* : if 50 < Number of employees < 250 and 10 mln €<Total assets < 43 mln €
  4. *Large enterprise* : if Number of employees > 250 and Total assets > 43 mln €
- **ATECO** , defined in order to get a lower number of unique values from the attribute ATECO2007code. The attribute has been obtained by grouping the codes into 21 macro sections established by ISTAT and identified by the first pair of digits in the ATECO2007code feature.
- **status**: has been divided between failed and active companies; record about merged and de-merged were deleted since we are not able to label them in a meaningful way.

```
#ACTIVED OR FAILED
df$`active` = 0
df$`active`[df$`Legal status`== "Active"] = 1
df$`active`[df$`Legal status`== "Active (default of payments)"] = 1
df$`active`[df$`Legal status`== "Active (receivership)"] = 1

df$status = "active"
df$status[df$active == 0] = "failed"
```

1.3 Failure Prediction

In the last section of the report 5, the goal was to build different models to predict before when a company is in rick of failure. Using supervised learning we are going to train different models, parametric and not, in order to detect what are the companies at risk and how much probable a company will fail. With this purpose we built a scoring model and a rating model.

2 Question A1: Failed and Active companies in 2018

In the following we are going to analyze the distribution of failed and active companies in 2018 of size and age. Our goal is to find out if the distribution change for a specific company form or for a specific industry sector in the year. For this purpose, after plotting the distribution of the various samples, we are going to use the non parametric KS test in order to calculate the distances between them. This kind of test usually has better results with Gaussian distributions. Indeed, the distributions at hand are inverse Gaussian distributions as we can see from the graphs. Moreover, we found another confirmation by plotting the distributions through boxplot. Furthermore, we tried to carry out a Shapiro-test but, due to the size of the dataset we couldn't obtain valuable results. In conclusion, we are going to consider only D (maximum distance) because p-value on a large sample loses consistency.

2.1 Legal Form

For the aim of investigating the distribution of active and failed companies based on the type of legal form, we used *ggplot* library to plot the age and the *standard R barplot* to plot the size [1], in combination with KS test for calculating the maximal distances between distributions of active and failed companies. The kind of legal forms that have been investigated are 9 (Consortium, S.A.S., S.R.L., S.P.A., S.N.C., etc.) and one of them (*"Others"*) groups those that have a relatively small sample.

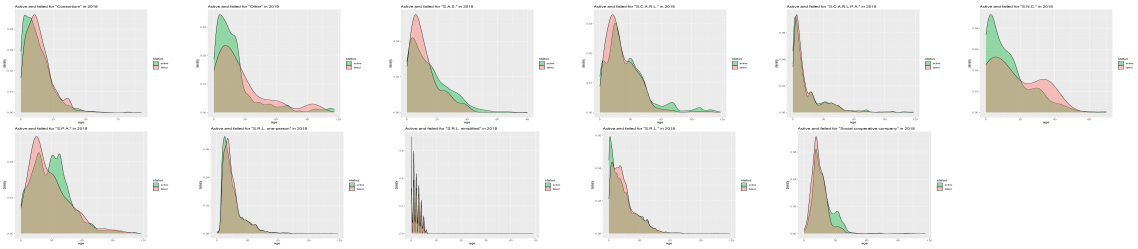


Figure 2.1: Age of Failed/Active Companies by Legal Forms

	Consortium	S.R.L.	S.P.A.	S.R.L. one-person	S.C.A.R.L. P.A.	S.R.L. simplified	S.C.A.R.L.	Social cooperative company	S.N.C.	S.A.S.	Other
D	0.099	0.0801	0.1117	0.0579	0.0489	0.0834	0.0887	0.1586	0.1866	0.1586	0.1841
p-value	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7771	0.5991	0.7054

Table 2.1: KS test (active/failed over age by legal form

2.1.1 Age

Looking closer at the results for *AGE*, see that the distributions having the major statistical difference are the ones of S.N.C., S.A.S. and Other. In particular:

- **S.N.C.**: during 2018, the distribution is mainly represented by companies having up to 20 years of age; those that falls into this range are predominately active whereas those that are more long-term (from 20y on) are predominately failed. This means that the probability of having an active company in the range 0-20 is higher than the one of a failed business. Inversely, companies from 20 to 55/60 years of age are more probable to be failed. Noteworthy, we can see that the distribution of failed companies has two peaks at around 5/6 years and 34/35 years.

[1] sinze it is an ordered qualitative variable, it was better to use a discrete visualization too, and not a contiguous one

- **S.A.S.**: during 2018, the distribution is mainly represented by companies having up to 15 years of age (relatively new companies); moreover, those that falls into this range are predominately failed whereas those that are older are for the most part active. We can clearly see that S.A.S. companies have an high probability to fail at an early stage (0-15 years).
- **Other**: during 2018, the distribution in mainly represented by companies having up to 20/25 years of age; furthermore, those that falls into this range are predominately active whereas those that are older are mostly failed. However, companies with more than 105/105 years of history are for the most part active. Both distributions have a peak at around 10 years and then they start decreasing. We can see from the graph that the probability distribution of active companies decreases faster than the one of failed business.

For the other legal forms, the distributions between active and failed don't have a significantly statistical difference. In fact, the two distributions have a relatively similar pattern (trend), thus having a very little D value in the KS test.

2.1.2 Size

Regarding *SIZE*, the majority of companies in the various samples is micro enterprise and active, in line with the Italian economic framework. In particular, by looking at the barplots we can see a clear similarity between different legal forms such as S.N.C., S.A.S. and S.R.L. Simplified. Looking at the S.P.A. plot instead (it has a clear similarity with Social Cooperative Company), we can find differences compared to the legal form S.N.C.:

- first of all, the distribution of active companies with respect to *SIZE* is uniform, in contrast with the one of S.N.C.
- for micro-enterprise there is a reversal trend respect to S.N.C. since S.P.A. has much more failed companies with *SIZE* 0

	Consortium	**S.R.L.**	**S.P.A.**	**S.R.L. one-person**	**S.C.A.R.L P.A**	**S.R.L. simplified**	**S.C.A. R.L.**	**Social cooperative**	**S.N.C.**	**S.A.S.**	**Other**
D	0.0888	0.1244	0.5381	0.1422	0.0855	0.0043	0.1561	0.3476	0.0349	0.1391	0.2285
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.9999	0.0000	0.0000	1.0000	0.7563	0.4309

Table 2.2: KS test (active/failed) over size by legal form

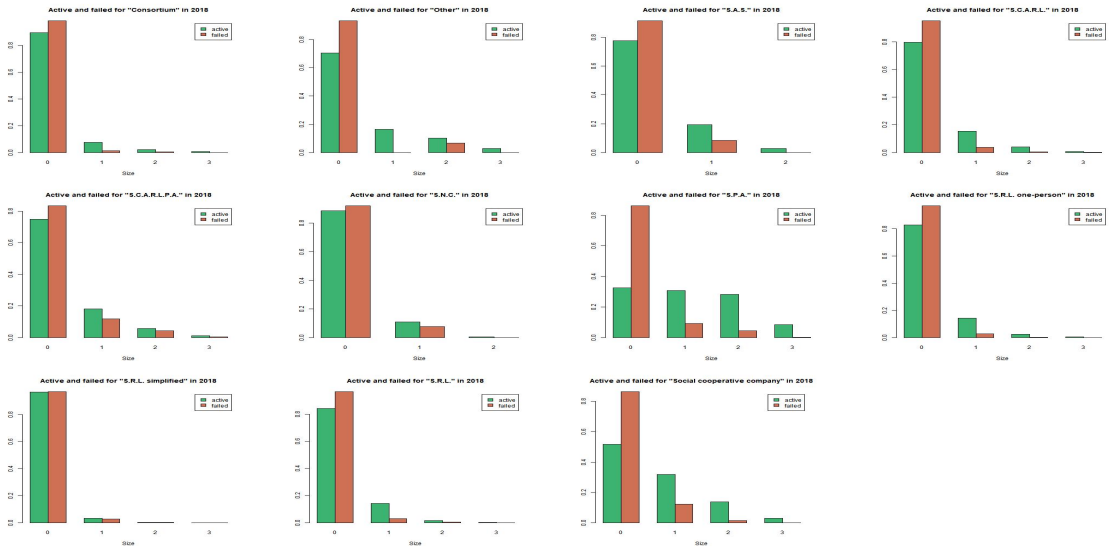


Figure 2.2: Size of Failed/Active)Companies by Legal Forms

2.2 ATECO sectors

For the aim of investigating the distribution of active and failed companies based on the type of ATECO, we started by grouping the codes in ATECO2007code feature from the first two digits, according to the ISTAT tree structure. We thus obtained 21 macro sections representing the various macro sectors.

2.2.1 Age

In the following, we are going to analyze the distributions having the main statistical significance in *AGE*.

In particular, regarding the section C (*Manifattura*) we can see that the D KS test is 0,3057 (the highest). This might be because of the sample size that is much larger than any other sections, thus reflecting the Italian economic framework. Looking closer at the graph, the two distributions are mainly represented by companies in their first 25 years of age. The main difference among the two distributions is in the first 20 years of age. In fact, the active ones are predominant at the very early stage (up to 5 years) while it is more probable to have failed businesses from 5 to 20 years of age. After that, we can clearly see a similar pattern between each other, although there is a slight increase in active companies from 20 to 47 years of age.

Regarding the section B (*Estrazione Minerali*), we can see that the D KS test is 0,228. Moreover, the two distributions are more uniform compared to the ones of the other sectors and the shapes reflect a Gaussian distribution. There is also a peak in the failed companies at around 18 years followed by a second peak in the "active" distribution after 25. The pattern is similar after 40y.

Regarding the section E (*Gestione Fogne e Rifiuti*), D KS test is 0,2573. Active companies are mostly at an early stage (in the first 10 years of age) and the distribution has a peak at around 5 years. As far as the failed distribution concern, we can see that it has two peaks: one at around 7/8 years and the other after 15 years. It means that we have a higher probability of having active companies in the first 10 years of age while failed ones predominate from 10 to 25.

	(G)	(C)	(H)	(E)	(F)	(K)	(Q)	(A)	(I)	(N)	(J)	(B)	(R)	(M)	(D)	(S)	(L)	(P)	(U)	(O)
D	0.0551	0.041	0.0673	0.0747	0.168	0.1597	0.1178	0.0862	0.0771	0.0666	0.0516	0.1028	0.1039	0.0901	0.0726	0.0689	0.1012	0.0470	0.6537	0.2078
p-value	0.0000	0.000	0.0000	0.0547	0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1728	0.0000	0.0000	0.0216	0.0005	0.0000	0.1693	0.0144	0.9644

Table 2.3: KS test (active/failed over age by ateco code

2.2.2 Size

Concerning the *SIZE*, we can clearly see that companies are mainly micro-enterprise regardless the industrial sector, especially for category (*immobiliari*) where companies are classified as micro for the very large part. Moreover, the slight predominance of failed companies for *SIZE=0* is present in every section, except for section T and U (due to the sample size). In particular, manufacture sector presents an interesting distribution. In fact, it has a predominance of micro-enterprise with the majority of the failed. However, we can see that there is a clear predominance of active companies for small, medium and large size enterprise.

	(G)	(C)	(H)	(E)	(F)	(K)	(Q)	(A)	(I)	(N)	(J)	(B)	(R)	(M)	(D)	(S)	(L)	(P)	(U)	(O)
D	0.1093	0.3057	0.1317	0.2573	0.0913	0.0641	0.2041	0.1031	0.1495	0.1015	0.1136	0.228	0.0835	0.0677	0.0452	0.0871	0.0032	0.118	0.0	0.1212
p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.000	0.0000	0.0000	0.3451	0.0000	1.0000	0.000	1.0	1.0000

Table 2.4: KS test (active/failed over size by ateco code

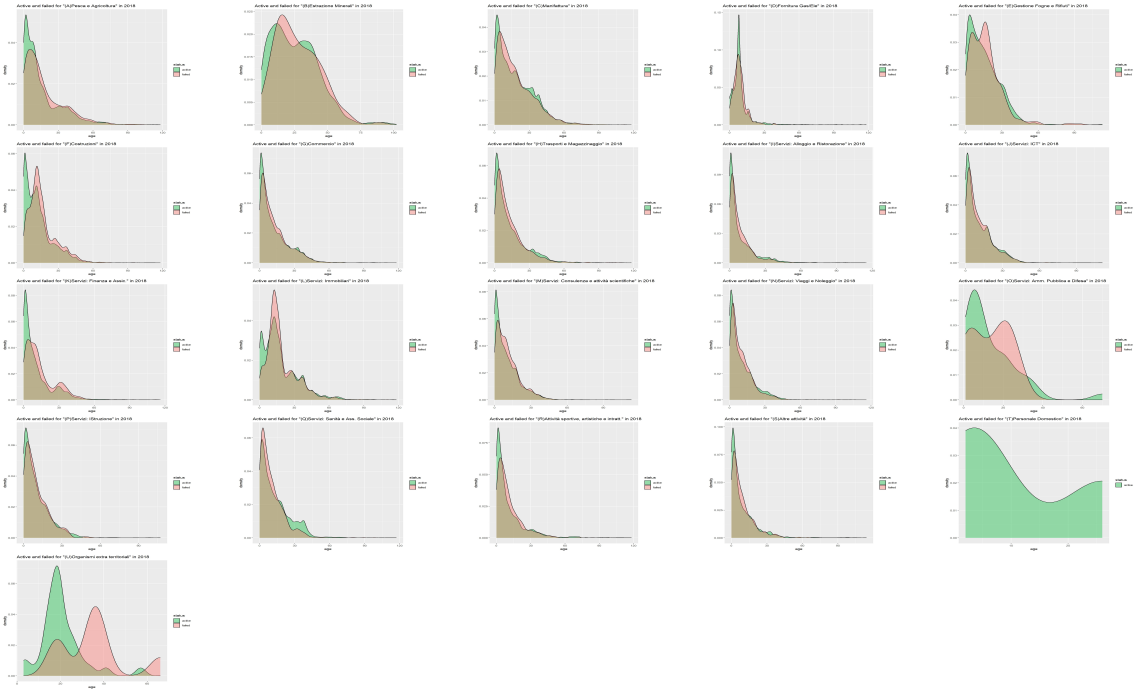


Figure 2.3: Age of Failed/Active companies by ATECO code

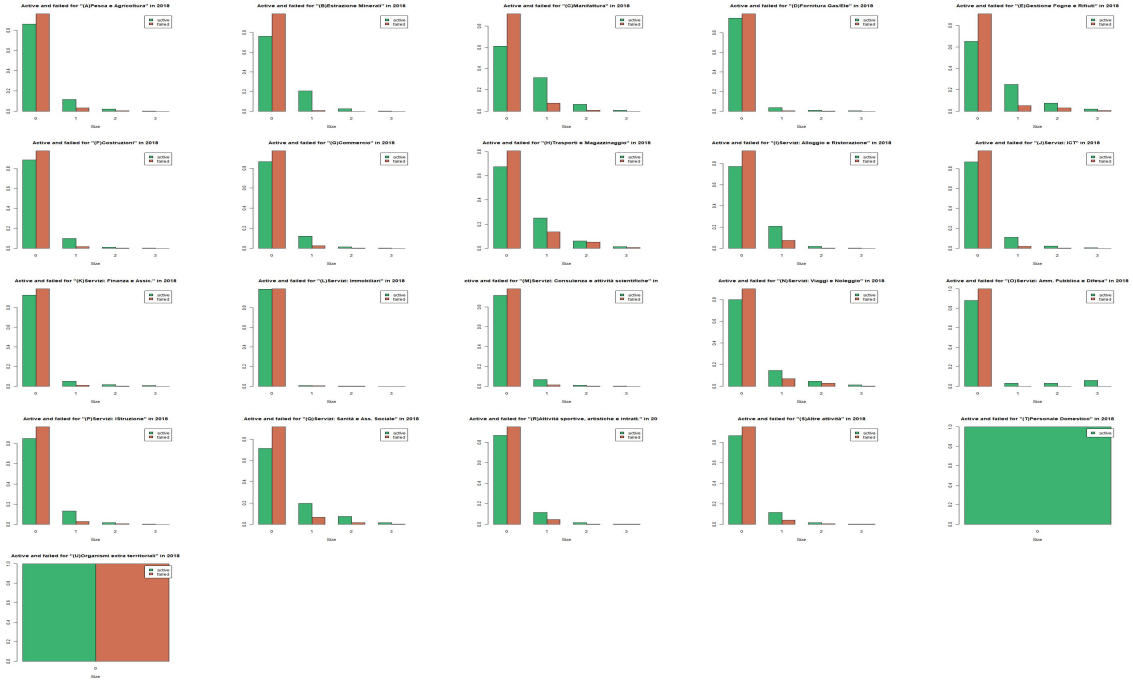


Figure 2.4: Size of Failed/Active by companies ATECO code

3 Question A2: failed companies over time

For the aim of investigating the distribution of failed companies over time based on the legal form, we decided to analyze them by **last 5 years** and by **period** (*1990-2010* / *2011-2014* / *2015-2020*).

3.1 Legal Form

3.1.1 Age

Looking closer at the graphs, what it turns out is that the distribution in the last 5 years of *AGE* is quite similar over time, with a higher density of failed companies in the early years of life. We can also see that, regardless the type of legal form, companies tend to fail less in recent years (2017-2018) compared to the previous ones in which the density of failed companies it higher, especially in the early stages. For instance, if we look at *"Consortium"* we can see that the peak in the distribution is slightly moving on over time, while the density in the first years of age decrease year by year, becoming pretty flat from 50. One of the causes that we can point out is the progressive recovery from the great recession that massively hit the national economy.

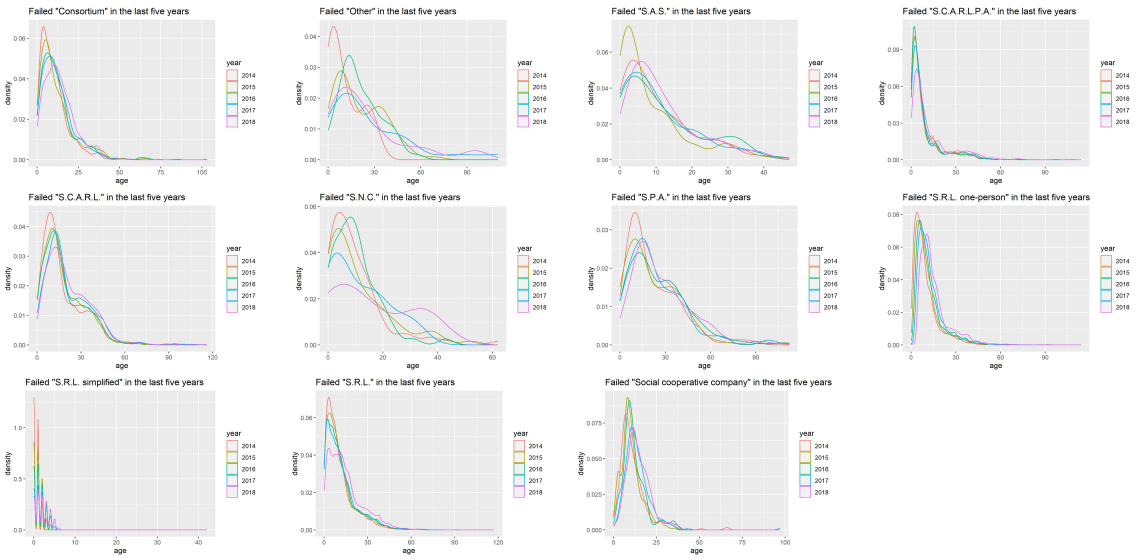


Figure 3.1: Age of Failed Companies by Year and by Legal Forms

	S.P.A.	S.R.L.	S.R.L. one-person	S.C.A.R.L.P.A.	S.C.A.R.L.	Consortium	Social cooperative company	S.R.L. simplified	S.A.S.	S.N.C.	Other
2014	0.0712	0.0164	0.0268	0.0384	0.0511	0.0274	0.1757	0.3078	0.0525	0.0570	0.2902
2015	0.0797	0.0351	0.0988	0.0355	0.1052	0.0452	0.1871	0.1635	0.1329	0.0557	0.1491
2016	0.1268	0.0558	0.1813	0.0530	0.1035	0.0939	0.2734	0.0662	0.0954	0.1463	0.2654
2017	0.1261	0.0776	0.2521	0.0390	0.1895	0.1075	0.3563	0.0804	0.0569	0.1791	0.2479
2018	0.1936	0.1968	0.3601	0.1241	0.1941	0.2017	0.3829	0.1552	0.1846	0.2572	0.2050

Table 3.1: KS test: failed over age by legal form and year

Regarding the distribution of failed companies by period, we can see that failed companies are predominately present in 1990-2010 and in first years of age. However, some of legal forms have interesting patterns such S.A.S. and S.C.A.R.L.P.A. in which the densities of failed companies pretty much match each other over the periods.

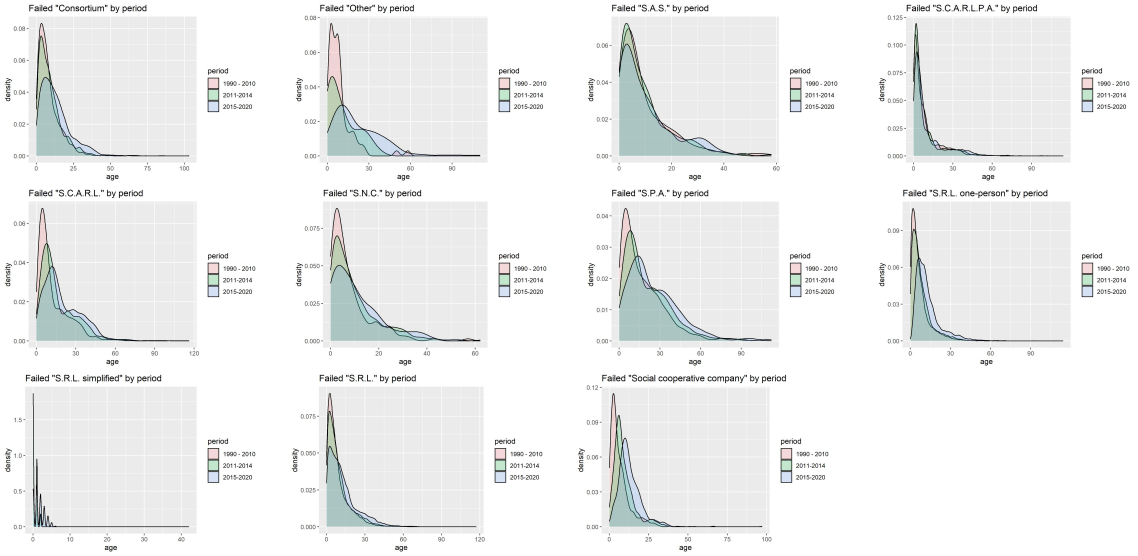


Figure 3.2: Age of Failed Companies by Period and Legal Forms

	S.P.A.	S.R.L.	S.R.L. one-person	S.C.A.R.L.P.A.	S.C.A.R.L.	Consortium	Social cooperative company	S.R.L. simplified	S.A.S.	S.N.C.	Other
1990 - 2010	0.0819	0.0022	0.0104	0.0604	0.0157	0.0027	0.0296	0.0309	0.0348	0.0232	0.1138
2011 - 2014	0.0112	0.0072	0.0091	0.0277	0.0090	0.0020	0.0078	0.0036	0.0074	0.0063	0.0347
2015 - 2020	0.1530	0.0086	0.0247	0.0099	0.0164	0.0005	0.0203	0.0005	0.0063	0.0120	0.0248

Table 3.2: KS test: failed over size by legal form and period

3.1.2 Size

Regarding *SIZE*, what it turns out is that the density of failed companies for micro-enterprise is higher in the 2015-2020 period regardless the legal form, except for S.R.L. simplified. For size 1-2-3 we can notice the predominance of failed companies in the 1990-2010 period, thereby with an inverse trend. To conclude, we didn't notice particular differences in the distribution for *SIZE* except for S.P.A., in particular in the last period; we evinced from the table 3.4 that is the most distant distribution from the general one.

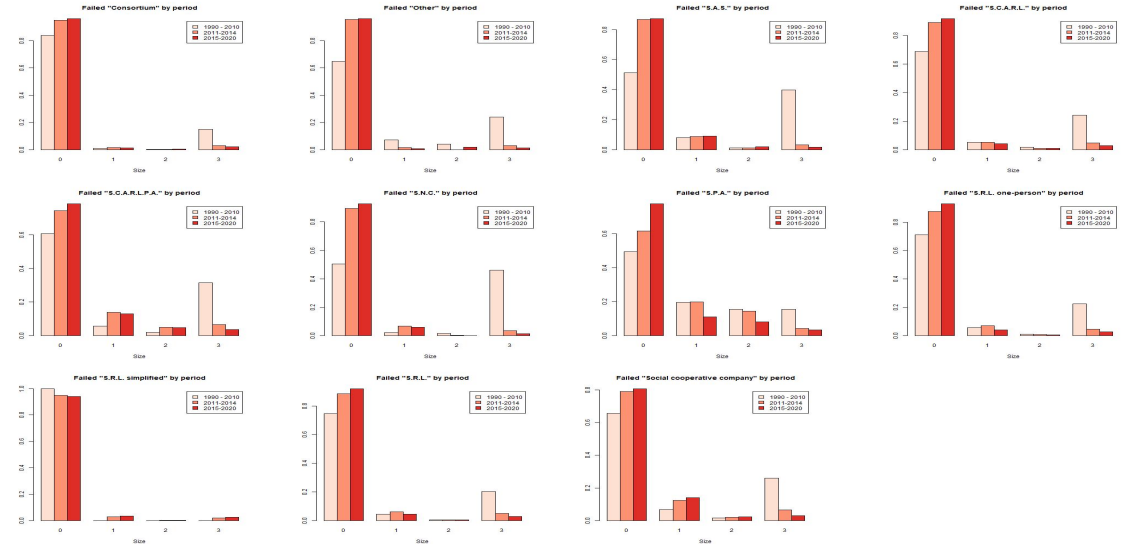


Figure 3.3: Size of Failed Companies by Legal Forms

	S.P.A.	S.R.L.	S.R.L. one-person	S.C.A.R.L.P.A.	S.C.A.R.L.	Consortium	Social cooperative company	S.R.L. simplified	S.A.S.	S.N.C.	Other
2014	0.0712	0.0164	0.0268	0.0384	0.0511	0.0274	0.1757	0.3078	0.0525	0.0570	0.2902
2015	0.0797	0.0351	0.0988	0.0355	0.1052	0.0452	0.1871	0.1635	0.1329	0.0557	0.1491
2016	0.1268	0.0558	0.1813	0.0530	0.1035	0.0939	0.2734	0.0662	0.0954	0.1463	0.2654
2017	0.1261	0.0776	0.2521	0.0390	0.1895	0.1075	0.3563	0.0804	0.0569	0.1791	0.2479
2018	0.1936	0.1968	0.3601	0.1241	0.1941	0.2017	0.3829	0.1552	0.1846	0.2572	0.2050

Table 3.3: KS test: failed over age by legal form and year

	S.P.A.	S.R.L.	S.R.L. one-person	S.C.A.R.L.P.A.	S.C.A.R.L.	Consortium	Social cooperative company	S.R.L. simplified	S.A.S.	S.N.C.	Other
1990 - 2010	0.0819	0.0022	0.0104	0.0604	0.0157	0.0027	0.0296	0.0309	0.0348	0.0232	0.1138
2011 - 2014	0.0112	0.0072	0.0091	0.0277	0.0090	0.0020	0.0078	0.0036	0.0074	0.0063	0.0347
2015 - 2020	0.1530	0.0086	0.0247	0.0099	0.0164	0.0005	0.0203	0.0005	0.0063	0.0120	0.0248

Table 3.4: KS test: failed over size by legal form and period

3.2 Company Location

3.2.1 Age

With regard to the investigation of possible changes in the distribution of failed companies for a specific location based on *AGE*, we plotted them by year and by period for every region (see figures in the appendix6.2). Although there is a slight difference in the density of failed businesses in the north comparing to the ones of the south (denser), the distributions are pretty much the same, without any noteworthy difference neither by year nor by period. The number of failed companies decreases year after year regardless the location, especially in the early years of age. The only fact we deduce by KS test is that distributions in 2018 are similar between them but quiet different from the general one.

	Sic	Pie	Mar	Val	Tos	Cam	Pug	Bas	Ven	Lom	Emi	Tre	Sar	Mol	Cal	Abr	Laz	Lig	Fri	Umb
2014	0.0377	0.0188	0.0130	0.1174	0.0219	0.0155	0.0257	0.0612	0.0260	0.0226	0.0134	0.0180	0.0261	0.0569	0.0184	0.0419	0.0235	0.0380	0.0137	0.0327
2015	0.0207	0.0288	0.0361	0.0523	0.0296	0.0185	0.0172	0.0545	0.0320	0.0310	0.0339	0.0448	0.0327	0.0412	0.0274	0.0307	0.0255	0.0159	0.0277	0.0273
2016	0.0243	0.0287	0.0276	0.0927	0.0269	0.0396	0.0286	0.0383	0.0601	0.0457	0.0412	0.0272	0.0265	0.0393	0.0410	0.0332	0.0319	0.0497	0.0420	0.0392
2017	0.0465	0.0368	0.0639	0.1584	0.0546	0.0433	0.0308	0.0565	0.0574	0.0506	0.0602	0.0593	0.0674	0.0618	0.0555	0.0551	0.0350	0.0609	0.0420	0.0475
2018	0.1651	0.1442	0.1983	0.2021	0.1640	0.1700	0.1568	0.1346	0.1696	0.1674	0.1691	0.1538	0.1966	0.1715	0.2023	0.1749	0.1454	0.1722	0.1805	0.1723

Table 3.5: KS test failed over age by region

	Sic	Pie	Mar	Val	Tos	Cam	Pug	Bas	Ven	Lom	Emi	Tre	Sar	Mol	Cal	Abr	Laz	Lig	Fri	Umb
1990 - 2010	0.0414	0.0501	0.0804	0.0973	0.0719	0.0546	0.0444	0.0367	0.0827	0.0668	0.0810	0.0705	0.0539	0.0555	0.0695	0.0505	0.0432	0.0708	0.0750	0.0670
2011 - 2014	0.0414	0.0270	0.0369	0.0396	0.0354	0.0274	0.0346	0.0460	0.0268	0.0337	0.0284	0.0458	0.0398	0.0417	0.0365	0.0490	0.0323	0.0334	0.0240	0.0342
2015 - 2020	0.0651	0.0694	0.0990	0.0986	0.0828	0.0659	0.0649	0.0567	0.0933	0.0869	0.0917	0.0879	0.0891	0.0613	0.0865	0.0813	0.0624	0.0902	0.0876	0.0812

Table 3.6: KS test failed over periods by region

3.2.2 Size

For *SIZE*, we analyzed the distribution exclusively by period. What it turns out is that there is no significant change for a specific location. Indeed, the distributions are pretty much the same with a predominance of failed micro-enterprise in the 2015-2020 period, while for large enterprise the major density falls into the 1990-2010 period. Statistically too, no evidence is observable.

	Sic	Pie	Mar	Val	Tos	Cam	Pug	Bas	Ven	Lom	Emi	Tre	Sar	Mol	Cal	Abr	Laz	Lig	Fri	Umb
1990 - 2010	0.0076	0.0161	0.0074	0.0380	0.0015	0.0022	0.0129	0.0440	0.0066	0.0014	0.0032	0.0116	0.0100	0.0100	0.0236	0.0224	0.0047	0.0032	0.0143	0.0088
2011 - 2014	0.0027	0.0020	0.0174	0.0037	0.0109	0.0043	0.0016	0.0012	0.0118	0.0149	0.0121	0.0100	0.0034	0.0079	0.0019	0.0018	0.0076	0.0171	0.0217	0.0162
2015 - 2020	0.0087	0.0169	0.0220	0.0360	0.0098	0.0056	0.0115	0.0350	0.0163	0.0115	0.0126	0.0166	0.0093	0.0114	0.0220	0.0197	0.0031	0.0142	0.0355	0.0215

Table 3.7: KS test: failed over size by region and period

4 Question B: Distribution of failures wrt size/age/other in 2018

In order to answer to the question B asking for the analysis of changes in the distribution probability of failures for Legal Form, industry sector and specific location in a specific year (2018), we processed the data in the same way as we did for question A since the attributes were the same. The only difference is in the inclusion of SAS and SNC in "Others" due to the sample size. The analysis was carried out on size and age. In the following, we're going to have a closer look at the graphs describing the distributions.

4.1 Legal Form

4.1.1 Age

In the following graphs (4.1) we show the Conditional probability density plots for each legal form according to their age. It turns out that, although there is a predominance of active companies along the density distribution, in the first years companies are mostly active while up to around 20 years of age they tend to be failed more. We can clearly see that there is a growing trend of failed companies in the first 20/25 years of age. The overall trend after this first stage is descending with picks and downs along the age.
Looking closer at the most interesting graphs, we can see that Consortium have a first peak in failed probability at around 35/40 years of age and then decrease again. From 60 years up to 75 companies tend to not fail, while after 75 there is a new peak in the density with probability up to 100% of failing.
Regarding "Others" (it includes SAS and SNC), after a first peak in the first 10/15 years, the trend decreases significantly. We have two new peaks over 0.50 at around 60 and 90 years. Despite up and downs in the distribution, the SRL legal form instead doesn't discord a lot from the probability 0.5 for every year of age.

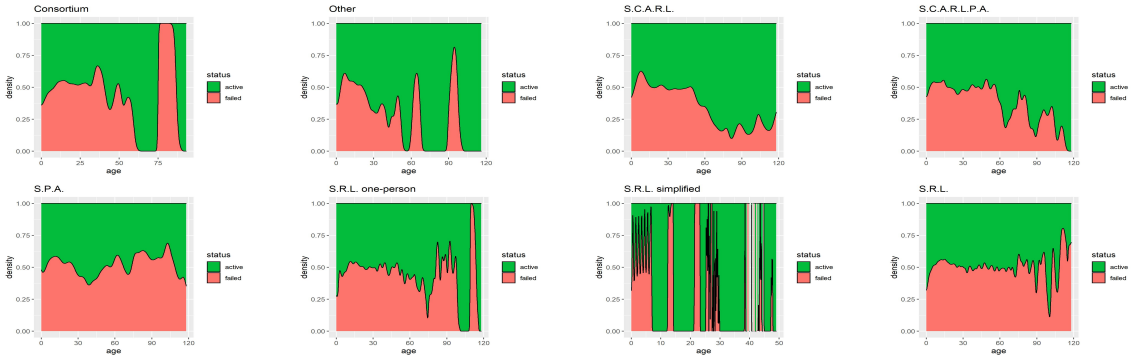


Figure 4.1: Age of Failed Companies by legal form

4.1.2 Size

Regarding the *SIZE*, we can see that there is an unbalance between active and failed in 2018. In fact, failed companies never go over 25% of density regardless the legal form. We can also see that the density of failed companies is higher for micro and large enterprises, except for S.R.L. simplified which presents a predominance of failed companies in medium enterprises comparing to the other sizes. Moreover, S.P.A. has the failed company count for sizes equal to

1,2,3 tending to zero, as we can see from its graph. Regarding the other legal forms, we haven't found any significance difference in the distribution of failures.

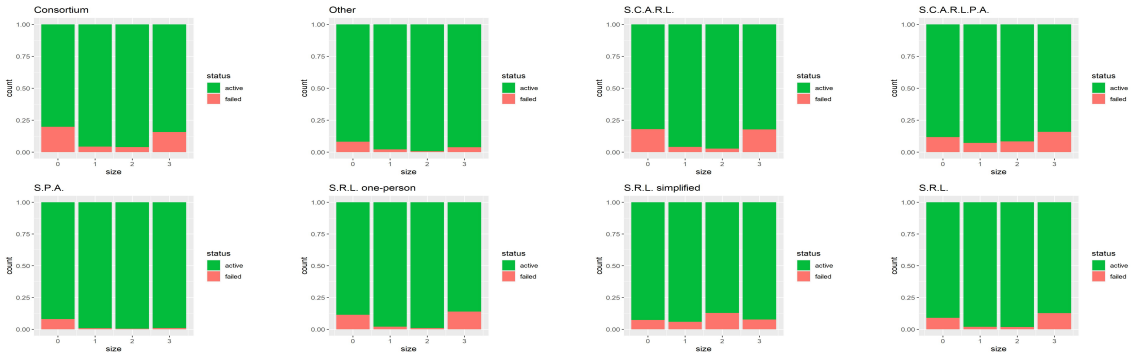


Figure 4.2: Size of Failed Companies by legal form

4.2 ATECO Codes

4.2.1 Age

The plots in figure 6.9 don't show any proper pattern or a particular characteristic trend. For sure the sample is more stable for the younger companies, since the sample is bigger , while the probability of failure become random (the distribution is oscillatory) with the age growth. This may be due to the sample size, smaller for older ages. Different Ateco sectors doesn't show failed company in the last ages, or maybe doesn't have older companies at all. The most extreme case is the *"Gas"* sector , that doesn't show failed companies older than 45 years old. It is more uniform the distribution in the three largest sectors ("A" ,"B","C"), where a good portion of the ancient companies had failed too , in particular in the *"manufacture"* the 75% of oldest factory failed.
The most interesting distribution is the one of "Extraterritorial Organization": in contrast with the other distributions, it shows younger company not failing too much, whereas almost all companies seem to fail after a certain age.

4.2.2 Size

Every sector distributions represented in figure 6.10 show that the percentage of failed company is really low for every age. However we can say for sure there is an higher probability of failure if the company is a micro-sized company or a large company. While between the PMI (small-medium companies) only a few are failing from what we can observe from the graphs. The highest(but still low) probabilities of failure, given the size small or medium, are for company of the sectors:

- Real Estate Services;
- Logistics and Warehouses;
- Travel & Rental

Instead in public administration, only the micro company show failures.

4.3 Company Location

In the following graphs (figures 4.3 and 4.4) we show the distribution of failures according to their age and their size in 2018. In particular, we're going to inspect possible changes in the distribution for a specific location. For this aim, we decided to categorize regions into five Areas: *Centro*, *Isole*, *Nord Est*, *Nord Ovest* and *Sud*.
For a more complete view of each region, please refer to the appendix 6.3 (figures 6.11 and 6.12).

4.3.1 Age

About the distribution over age, What is evident is that there is a slight descending trend for *Centro*, *Nord Est* and *Nord Ovest* where companies tend to fail less as age increases. Moreover, companies located in *Centro*, *Sud* and *Isole* have an higher probability of failure especially in the first 50 years where the distribution of the north area is under 50%, except for a peak in both of the distribution during the first 20 years where the trend of failures is growing. We can also see that the two areas with the highest probability of failures are *Isole* and *Sud*. In fact, the two distributions are pretty similar with a probability of failures that settles over 50% during the first 50/60 years, followed by up and downs with some peaks over 75%.

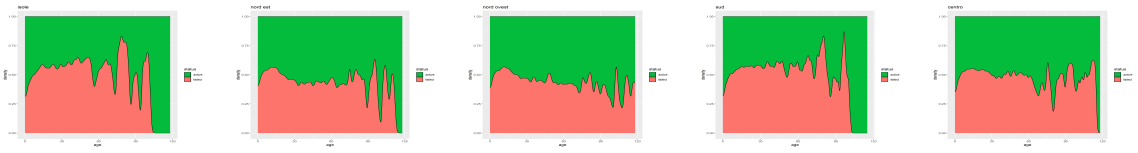


Figure 4.3: Age of Failed Companies by Geographical zone

4.3.2 Size

Regarding the SIZE, there is no significant statistical difference in the distribution of the Areas. If fact, what it turns out from the graphs is that regardless the location of the companies, the highest probability of failure is for micro and large enterprises whereas for small and medium enterprises the probability decreases significantly. Even looking at the graphs of each region there is no particular change in the distribution of failures.

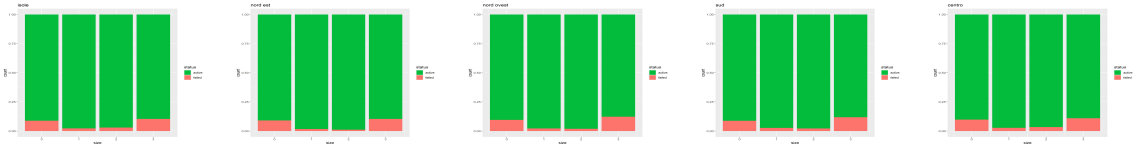


Figure 4.4: Size of Failed Companies by Geographical zone

5 Question C: Failure Prediction

In this section we describe how we build statistical models to **predict risk of a business failures** for the company of the *aida dataset*. The process start with cleaning the data and selecting some features using both **parametric and non parametric models**, trained by *cross validation technique*. After that, we split cleaned data into training and test set, in a temporal manner,with the objective to **use older data to predict newer**. The goal is to have both a *scoring model*, obtained by a regression and then to identify some ranges through a *rating model*. However we started performing a **binary classification** using the legal status of the company, labeled as *active* and *failed*.

5.1 Data Cleaning

Starting from the almost two million records dataset, we can proceed cleaning the data, just removing null values; in fact, they are enough to train a non parametric model, and so a parametric one. However, since we have ratios and indices of the last three years, we fill null record of two years ago with the values on three year ego, and so on for the record of the last year with the ones of two years ego. At this point we check the amount of null values by feature. What we observe is that three attributes in particular can influence the size of the data:

- Banks turnover
- Cost of debit
- Return on investment (ROI)

So we tried one thing: simply **deleting all null values**, lead us to *still have 200k records* in our database, most from 2018; while **deleting first the three featured cited above**, and then removing null values, allow us maintaining the *size of data over 500k*. In both cases, we have theoretically enough data to train a model, so during feature selection at 5.4, the importance of each feature will be discussed.

5.2 Introduction to Statistical Models

We decided to train both parametric and non parametric models. We used **linear discriminant analysis** in principle, for its importance in failure prediction, and for the ability in estimating feature importance that is one of our objective in this phase. The model is a parametric model and it has been evaluated using accuracy and cross validation. Another parametric model that we adopted is the logistic regression, useful method for binary classification and suitable for our case. The trained version is actually a **boosted logistic regression** that uses different samples from the dataset, in addition to the linear regression. The last model we chose is a machine learning ensemble technique, the **random forest**, suitable when you have a lot of attributes since this technique selects at each iteration different subsets of at attribute. It can be useful to evaluate features importance also for using further parametric models.

5.3 Model Requisites

Each one of the proposed models has got its own requisites.

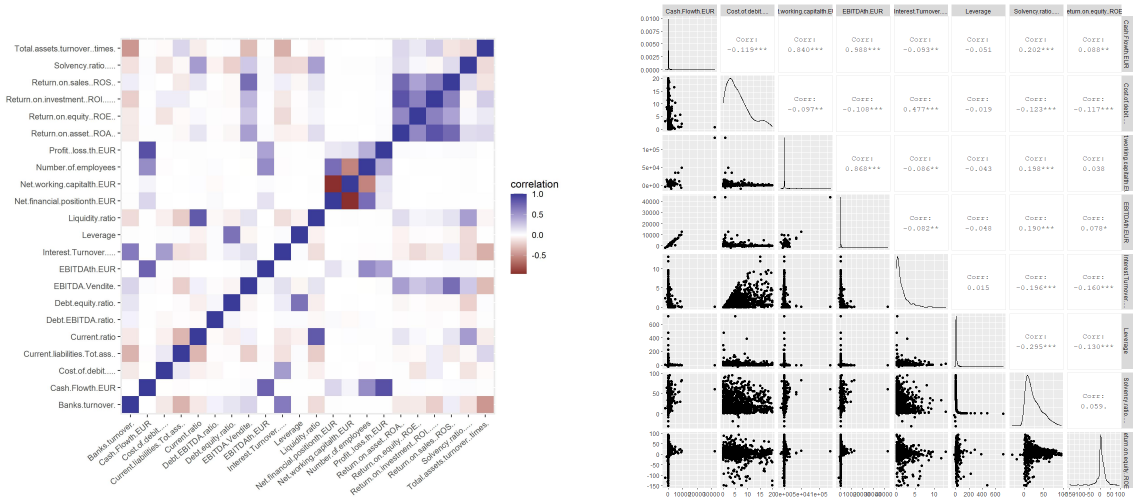
The most versatile between them is for sure the **random forest**, for which we do not have to select the features or check requirement, since it is usually strong enough to work with high correlated features and any type of numeric distribution. For this reason we will train it using the whole set of features.

Binary logistic regression requires the dependent variable to be binary and the observations to be independent of each other. Furthermore, **logistic regression** assumes linearity of independent variables and log odds; this does not require the dependent and independent variables to be related linearly. And finally requires that there is not multicollinearity among the independent variables.

Multicollinearity is a basic requirement also with regard to **LDA** and is variation. Moreover, before we try to fit the LDA model we must check for other prerequisites which are the multivariate normality and the homogeneity of the variance. We are going to discuss them before the application of the model.

5.4 Feature Selection

In this phase we use the dataset of 200k rows described in chapter 5.1. We exclude every categorical (region, province, ateco, legal form) examined in section 2, 3 and 4; and also the ordinal attribute we have (*size* and *age*) are excluded; so we concentrated on every ratios and indices that are traditionally used for bankruptcy prediction.



(a) Correlation Heat Map

(b) Scatter Matrix

Figure 5.1: Attributes Distributions and Correlations

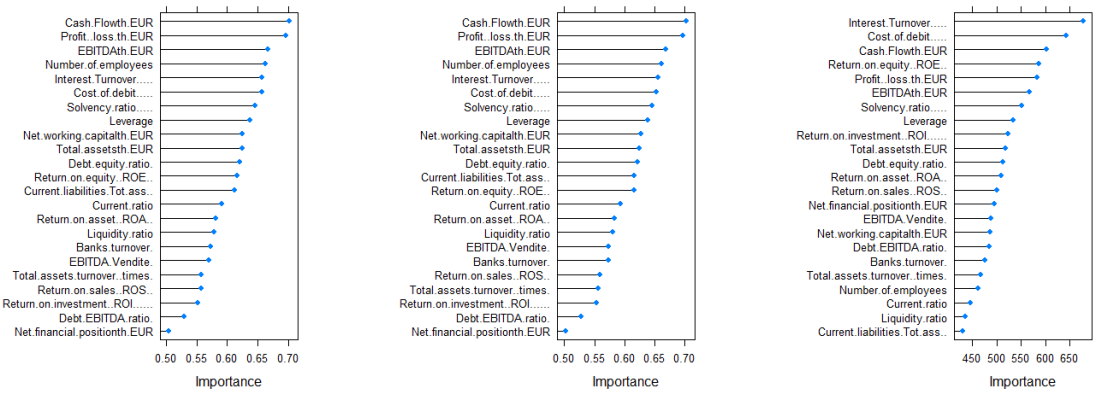
5.4.1 Correlation and Feature Importance

The first step to come across is to check correlation between attributes in order to remove redundant features. The easiest way is inspect the heat-map in figure 5.1a. Luckily, most of the features are uncorrelated and we can use them together through classifications algorithms. However, we can identify some correlation:

- ROS, **ROI**, *ROE*, ROA, *EBITDA Vendite*
- Net. Financial Position inversely correlated with Net. Working Capital
- cash flow turnover with profit loss and with EBITDA(€)
- liquidity ration with current ratio

The idea is to cross this analysis with the feature importance provided by the three models of section 5.2(trained by cross validation on the 200.000 rows and above all features), and decide what to keep and what to delete. In figure 5.2 it is possible to compare the importance of each

feature from model to model, with the features ordered from the most fundamental to the most redundant.



(a) LDA (b) Boosted Logit (c) Random Forest

Figure 5.2: Feature Importance by Model

What we observe immediately is that the feature importance given by LDA and logit is pretty the same, wheres random forest result is considerably different from the other two applied techniques. For this reason we are going to consider the two as a unique vote. After a double inspection on the heatmap of the the plots of figure 5.2, we made the following decisions :

- keep **ROE** because it is valued as the most important in the group of correlated ratios, while ROI, ROS (not really important) and ROA and ebitda/vendite are discarded
- Profit Loss and and **Cash Flow** are both really important, but since they have a Pearson correlation of 0.87 we only picked the most important, so the Cash Flow
- we take the **Net Working Capital** and not the Net financial Position
- we discarded the Banks Turnover, it is not important and contains a lot of null values; **Cost of Debit** contains a lot of null values too, despite this we should keep it for the importance attributed to it by all the three algorithms
- we take some important features: **leverage**, **interest turnover** , **EBITDA**, **solvency ratio**

Consequentially the the preferences expressed, the importance of the attribute *Cost of Debt* pushed us to go ahead with our models and evaluations using the dataset with 200k rows, because the bigger one, the one with 500k rows, doesn't contains the forementioned .

5.4.2 Normality, Variance Homogeneity, and Multicollinearity

On the selected attributes we advanced with further analysis. In order to perform Linear Discriminant analysis, as we mentioned before, we tried to verify the presence of some prerequisites that the datasetdistributions should satisfy:

- **Multivariate normal distribution**. We verified the absence of this requisites in different manners. Generally, in order to verify if a distribution is normal or not we can show the scatter plot and if the records are distributed among the centre of the axis we may have a confirmation of this property presence. Because of our normality check is multivariate, we had to verify if most of the scatter plots of the various attributes were respecting this distribution. This clearly form the 5.1b that is not respected. Furthermore we tried to have a confirmation by a measure that is able to give us a numeric result. In fact, we performed the Skewness and Kurtosis measure that give us the confirmation of the absence of this property.It is a measure of the tailedness of the probability distribution of a

real-valued random variable. From the result that we present in the next table it is clear that the various distributions are not normal. For most of them, despite the carried out normalization, the result are not satisfactory.

- **Variance homogeneity:** To test if the variance is homogeneous we performed the Bartlett test. We obtained a very low p-value. That was sufficient to confirm that there was not homogeneity of the variance.
- **Multicollinearity**. To check if this property/requisite holds among our distributions, we verify it in two different ways. Firstly, we obtained a scatter plot and a covariance matrix (heat map). From these two graphs we can't identify an high correlation between variables, so that it was a starting point to declare the absence of multicollinearity. Furthermore, we performed the *VIF* whose results are presented in the table below. A value of the Beta less than 10 suggests us that there is not multicollinearity among the different distributions. It happens for some of the attributes but we verified that deleting them does not change the LDA result too much. [2]

	LDA Coefficients	Logit Coefficients	Variance Inflaction Factor	Kurtosis	Skewness
Cash.Flowth.EUR	-0.0019316551	-1.32881	31.479180	48371.02	198.3723
Cost.of.debit.....	0.3323820630	0.28061	1.243198	3.393403	0.9876743
Net.working.capitalth.EUR	0.0001821493	-0.33089	24.776334	154649.9	-354.3376
EBITDAth.EUR	-0.0283896139	-1.49169	2.728653	46649.95	192.5706
Interest.Turnover.....	0.5424354918	0.31274	1.249769	12.84528	2.563149
Leverage	0.0993117810	0.05005	1.033500	6008.529	61.96045
Solvency.ratio.....	-0.3509125731	-0.36428	1.061304	3.757256	1.064316
Return.on.equity..ROE..	-0.4032024491	-0.23731	1.042555	10.93002	-1.590174

Table 5.1: Most Important Attributes examination

5.4.3 Models

We trained using on the 200k rows dataset with cross validation and accuracy to evaluate them. Performances are similar, accuracy is around 85% for LDA and Logit, with LDA a little stronger, and 0.867 for random forest. Kappa score instead is definitely higher for random forest, with an advantage of five percentage points (0.20 to 0.15 the others), however it denotes a poor concordance. The trouble in the trained models is that the accuracy around 0.85 seems to respect the prior probabilities of the two classes, that are in fact distributed in the manner: 85% of the total are active companies and only 15% of the companies are represented by the failed class. Actually we used also specificity (True Negative Rate) [3] to measure the model validity; what we obtained is very low performances(0.16 the LDA).Then we conclude that, despite the relatively high accuracy, the models are wrong and inefficient.

5.5 Temporal Split

In this section we explain how temporal information were used to split data into a training and a test set. Usually the good practise in model building, is to have a 70% of the data on the training set and the remaining part on the test. In this case, since we want to use older data to predict the newer, we can not split exactly randomly but we should choose a procedure. The first idea was to split data by the year from which the record is. So we decided to check the distribution over the year of the values, making it cumulative, to find the closest value to 70%. Data are plotted in a stacked bar chart (figure 5.3) that put in evidence also the distribution

[2] To perform VIF, we checked the ratio of the two variances using a logistic regression model.

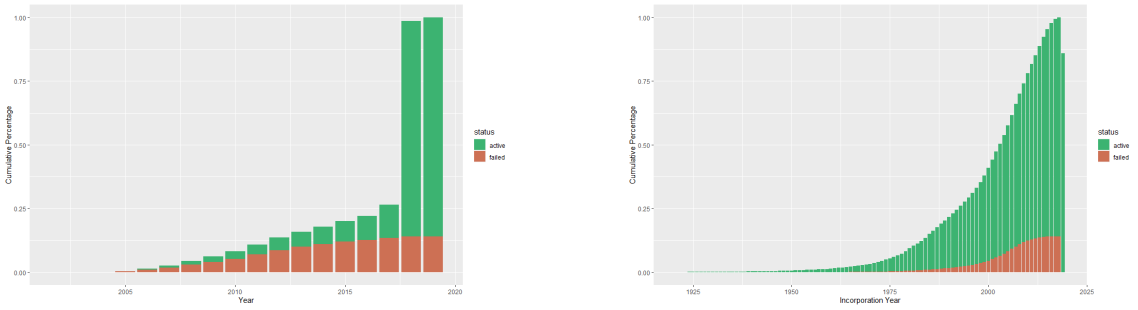
[3] In this case the specificity is the proportion of those failed companies who are correctly identified as failed

between the two possible values of the target attribute. We note a particular situation[4] on data and two main issues emerged by this analysis:

1. almost 75% of data are from 2018, and it makes impossible to use only the previous data (<= 2017) as test set to predict those;
2. while before 2018 the records are equally distributed between the two classes, in 2018/2019 the situation is different: around 99% of data are active companies;

The best approach to solve the first issue and partially even the second one was in our opinion to divide 2018's records into test and training set, in order to respect the traditional proportions. Moreover, we can impact positively on the imbalance of the test set, decreasing it [5].

However, we tried to think another possible temporal split, suitable both to do prediction and to respect the needs of the models. The best idea was to use the incorporation year... in this way we can use older companies to predict what is going to happen with the newer. In figure 5.3b it is possible to observe the cumulative distribution that we used to decide where to split. Using 2009 as threshold it is possible to have 70% of the data on the training set and 30% in the test; moreover, the two datasets have got the same balancing (85%-90% of companies are "active") between the positive and negative class.



(a) Year of the record

(b) Incorporation year of the company

Figure 5.3: Cumulative Distributions between active/failed

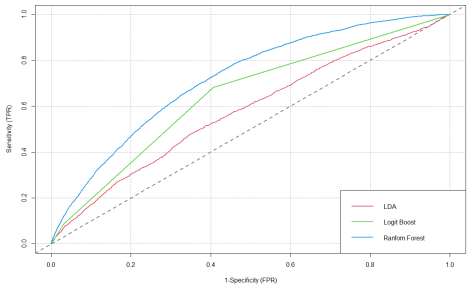
5.6 Training

At the end we split our data using the incorporation year: all companies incorporated before 2009 belongs to the training set; the remaining are the test set.

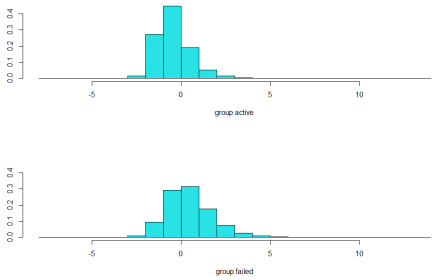
The training set is composed by 76020 observations, while the test set of 35825 ; so the split was made exactly at the 70%. However, the distribution of data among the two classes is not exactly the same, and it's still really unbalanced also in the training set. For this reason we decided to use weights to balance the decision in the models. The applied algorithms have been already introduced in section 5.2. This time we trained them using Cohen's Kappa as metric over cross validation. In fact, accuracy has been already declared as not reliable for the evaluation of these classifiers. It is actually better to look to the balanced accuracy, and to the specificity.
N.B. Only random forest was trained using all continuous attributes; while logit boost and LDA were trained using the features selected.

[4] Each company appears only once in the dataset, so older active company appears in the last year information about them were available, while older failed companies appear probably in the year the failed. This can justify the two issues encountered

[5] moving to the training set only active companies from 2018 is possible to make the distributions a little more similar



(a) ROC curves



(b) Classes Distribution on LDA

5.7 Evaluations

So the trained models have been evaluated with the test set. We were able to inspect the confusion matrix and to check all the metrics. We also used ROC curves (figure 5.4a) to compare the three models; by the inspection of this we know AUC is pretty higher than the others from random forest. Moreover, we can conclude by looking at the table 5.3 that it was the best fitted model, able to predict more failed companies respect to the others; with the highest Kappa and specificity it scored 0.55 in balanced accuracy. In the figure 5.5, we can examine all evaluation metrics plotted for this model using *"failed"* as positive class. AUC has a good score, covering 74.4% of the area. Precision is high only for the first records, then it starts to decrease, and we deduce the reason by looking at the distribution of the two classes over the probability. Recall and accuracy are more stable than the precision. The calibration shows that the number of failed companies is overestimated; having a slope less than one indicates optimism; the intercept instead is on 0.

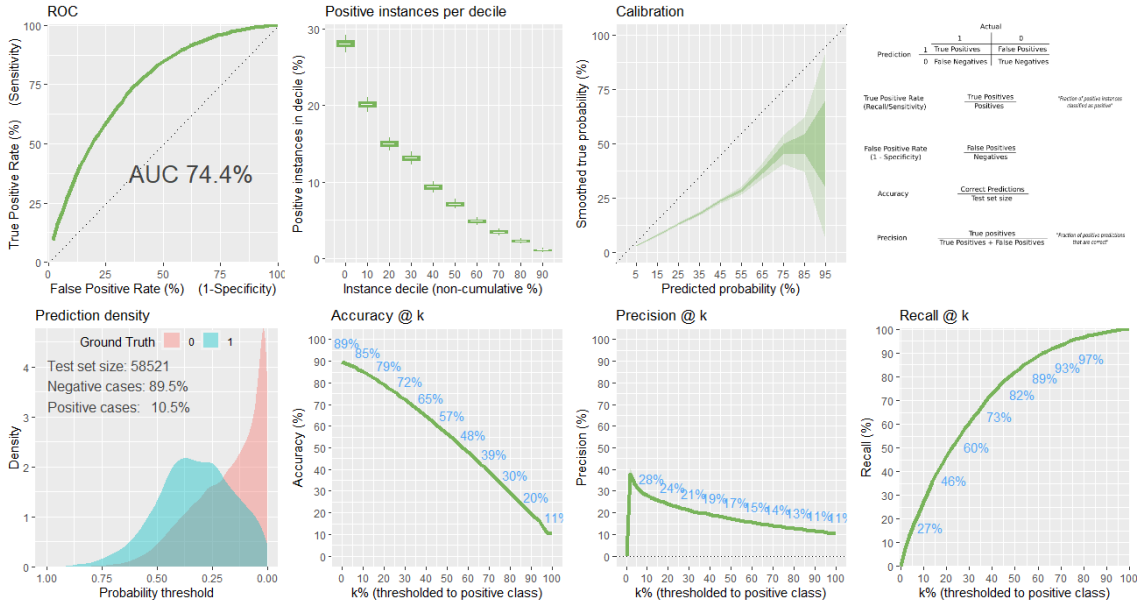


Figure 5.5: Random Forest Evaluations

5.8 Undersampling

Given the unbalance of the dataset, which can be a issue on the model performances; we tried a different temporal split, in order to have a bigger training set in which we could make random undersampling. Data are split between before 2012 and after. The undersampling is applied

on the training set, only in the active class. We choose to have the prior of the training set 1:2, so we had 50680 positive observations and 25340. Almost the 70% of the observation were in the training set at that point. So models are trained with the same parameters of before: using weights and cross validation with Kappa as metric.

		Reference	
		active	failed
Prediction	active	50413	5217
	failed	1961	930

(a) Random Forest

		Reference	
		active	failed
Prediction	active	23850	1237
	failed	9660	1078

(b) Logit Boost undersampling

Table 5.2: Confusion Matrices on the Temporal Split

Models are evaluated as before, comparing ROC curves and confusion matrix. ROC curves did not change very much respect to before, but from the table 5.3 it is visible how the undersampling impacted negatively on the random forest, while positively on the logit boost. In fact, both for the logit boost and for the LDA, the specificity has increased, while negative predicted value decreased a lot; but at the end, this is what we wanted: it is preferable in this case to have false negatives rather than false positives. However the logit boost with a specificity of 0.46 and a balanced accuracy of 0.59 seems to be the best model trained, also better than the random forest before the undersampling. In table 5.2 is reported its confusion matrix together with the one of the random forest, so we can compare them. Obviously it remember us there are less samples on the undersampling part, because of the different split. Then, we see more true negatives are guessed by using the logit boost, and the false positives are less than the middle with respect to the random forest. Moreover, evaluating a total metric, also the balanced accuracy is higher in the logit model.

	Models			***Undersampling***		
	LDA	**Logit Boost**	**Rando Forest**	**LDA**	**Logit Boost**	**Rando Forest**
Balanced Accuracy	0.52262	0.5328	0.5569	0.5342	0.5887	0.5407
Specificity	0.06198	0.1079	0.1513	0.1369	0.4657	0.1127
Neg. Predicted Value	0.30286	0.2308	0.3217	0.1212	0.1004	0.1986
Kappa	0.0697	0.0859	0.1486	0.0644	0.0659	0.1018

Table 5.3: Evaluation Metrics

5.9 Scoring Model

However, if we observe the logit evaluation plots of figure 5.6, we can note that AUC is lower, recall and accuracy (not balanced) are low, and other measurements are really confusing and shown the behaviour is not stable. Finally, classes have strange distribution with three different pick of probability; while in the random forest the distribution of probability make more sense. So if we want to choose a clear and performing scoring, that at least give us a sensitive probability... we would choose the random forest; while if all we want is a binary classification with an high specificity for the failed companies, we should adopt the logit boost model after the undersampling. Moreover by looking at the performances of the models on the self training set, we discovered logit boost performance are totally different from those expected, confirming the randomness of the test prediction good performances. In the opposite, the random forest

is able to predict the training set without any error; this makes it finally become with no doubts the favorite model.

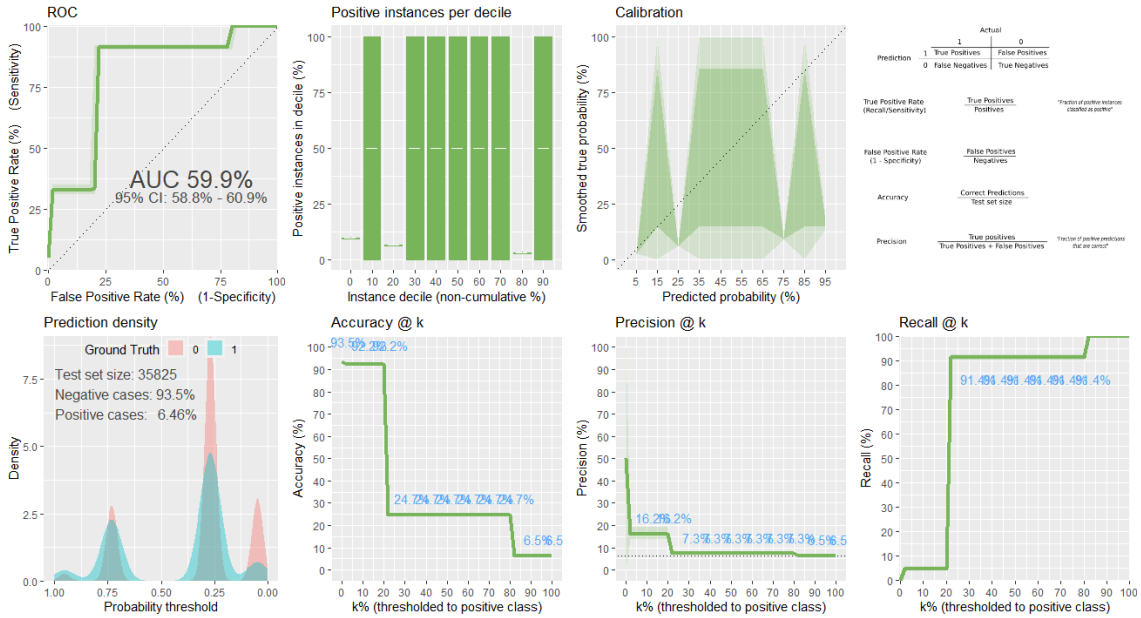
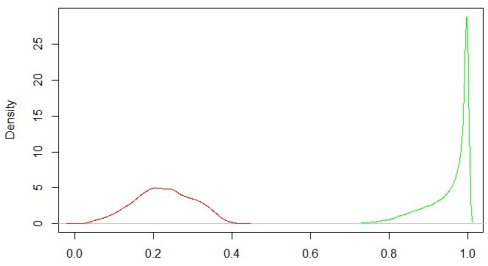


Figure 5.6: Logit Boost Undersampling Evaluations

5.10 Rating Model

The basic idea for the rating model is to create another class with a low risk of failure, like a neutral area between positive and negative, that can be useful to insert a portion of false negatives and false positives. Then, we can evaluate the result only on the original two classes. Since the distributions of prediction probabilities of the training set where totally separated (figure 5.7a), we can identify that gap as the portion of probability belonging to the neutral class; moreover we decide to look at the distributions over the test too. So our approach is aposteriori approach, that divide the sample in classes using the score. Precisely, records with a probability to be active, between 55% and 75% are labeled as *low risk of failure* or neutral; and the confusion matrix(5.7b) is shown over the original two classes, the *"not neutral"* records. True failed are always the same number as in table 5.2a; instead a lot of false active have been removed, but a lot of true active too. Summarizing, the specificity increased to 0.24 while the balanced accuracy grew up to 60%. The neutral class contains 14671 samples: the 25% of the total.



(a) Distributions on the training set

		Reference	
		active	failed
Prediction	active	38314	2643
	failed	1963	930

(b) Rating Model Confusion Matrix

Figure 5.7: Random Forest Rating Model

6 Appendix

6.1 Appendix A.1

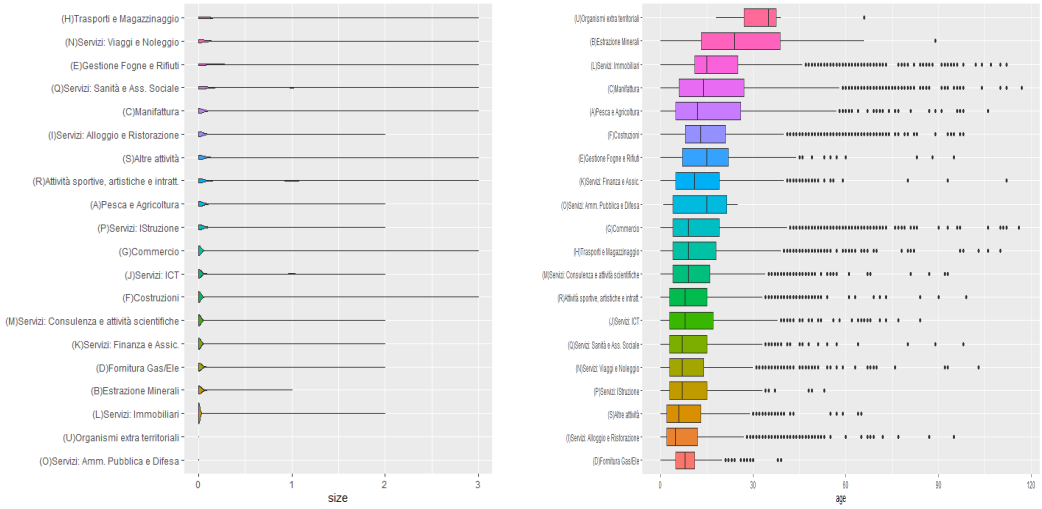


Figure 6.1: Size and Age of Failed Companies by Ateco code

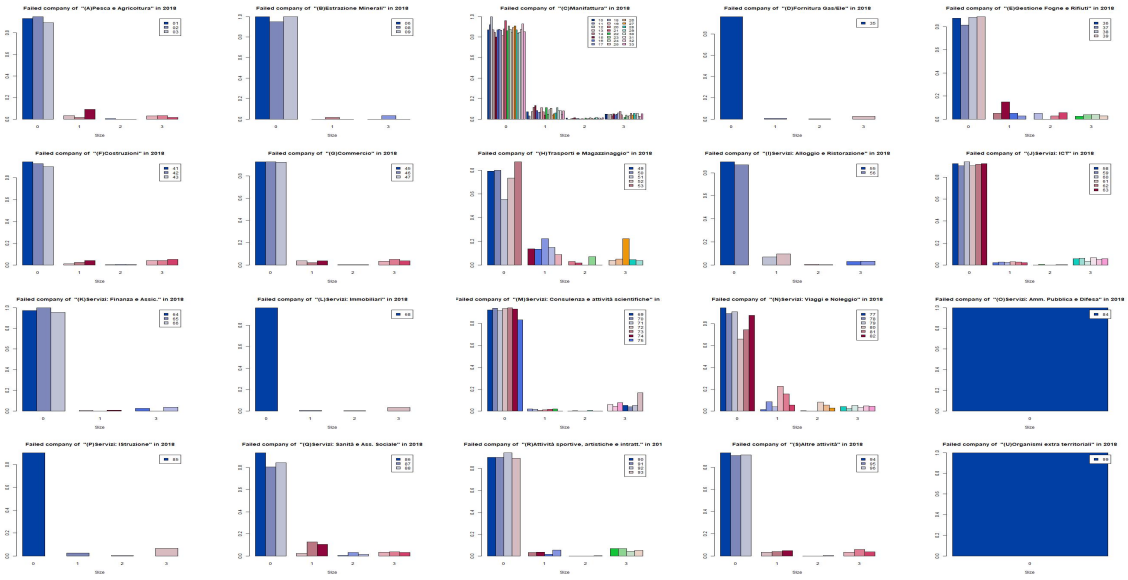


Figure 6.2: Size of Failed Companies by ATECO sectors

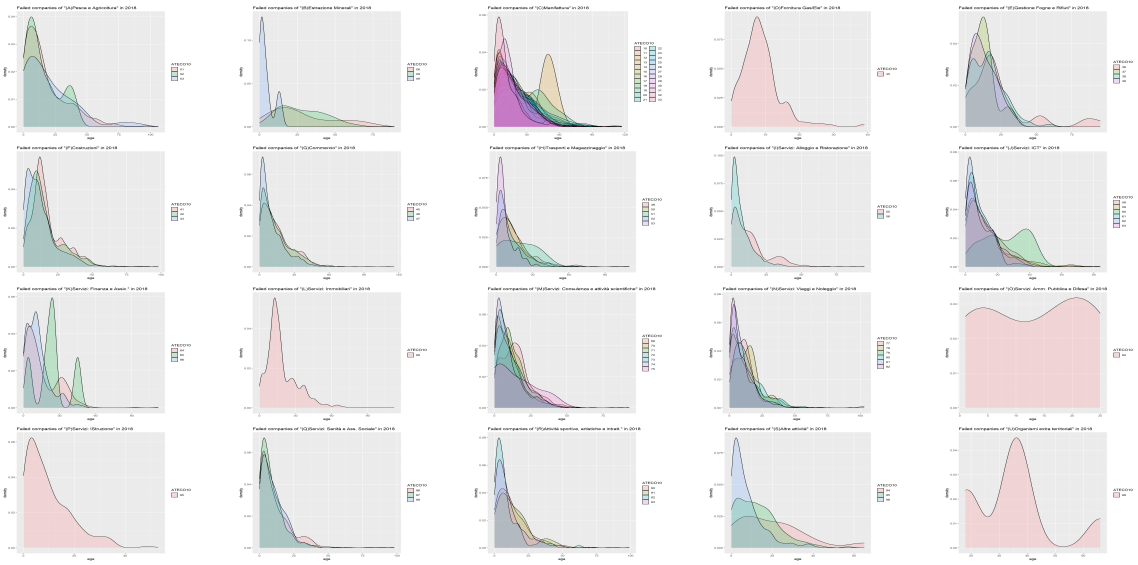


Figure 6.3: Age of Failed Companies by ateco sectors

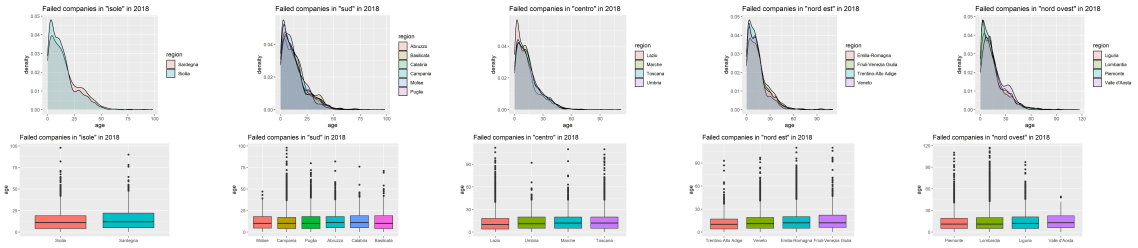


Figure 6.4: Age of Failed Companies by region

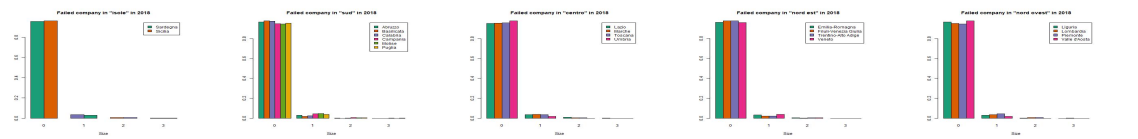


Figure 6.5: Size of Failed Companies by region

6.2 Appendix A.2

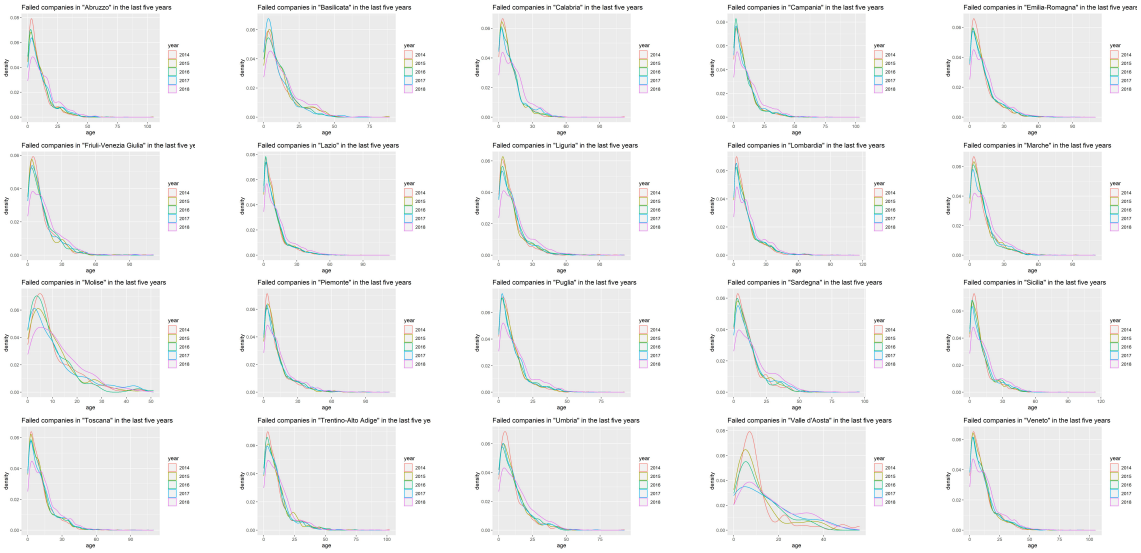


Figure 6.6: Age of Failed companies by year and by region

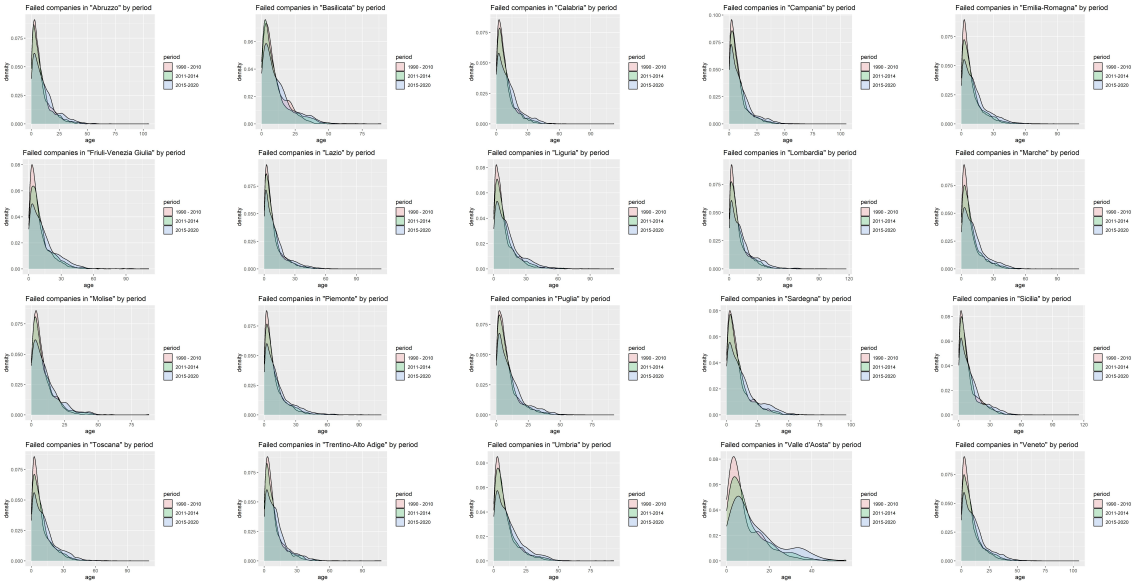


Figure 6.7: Age of Failed companies by period and region

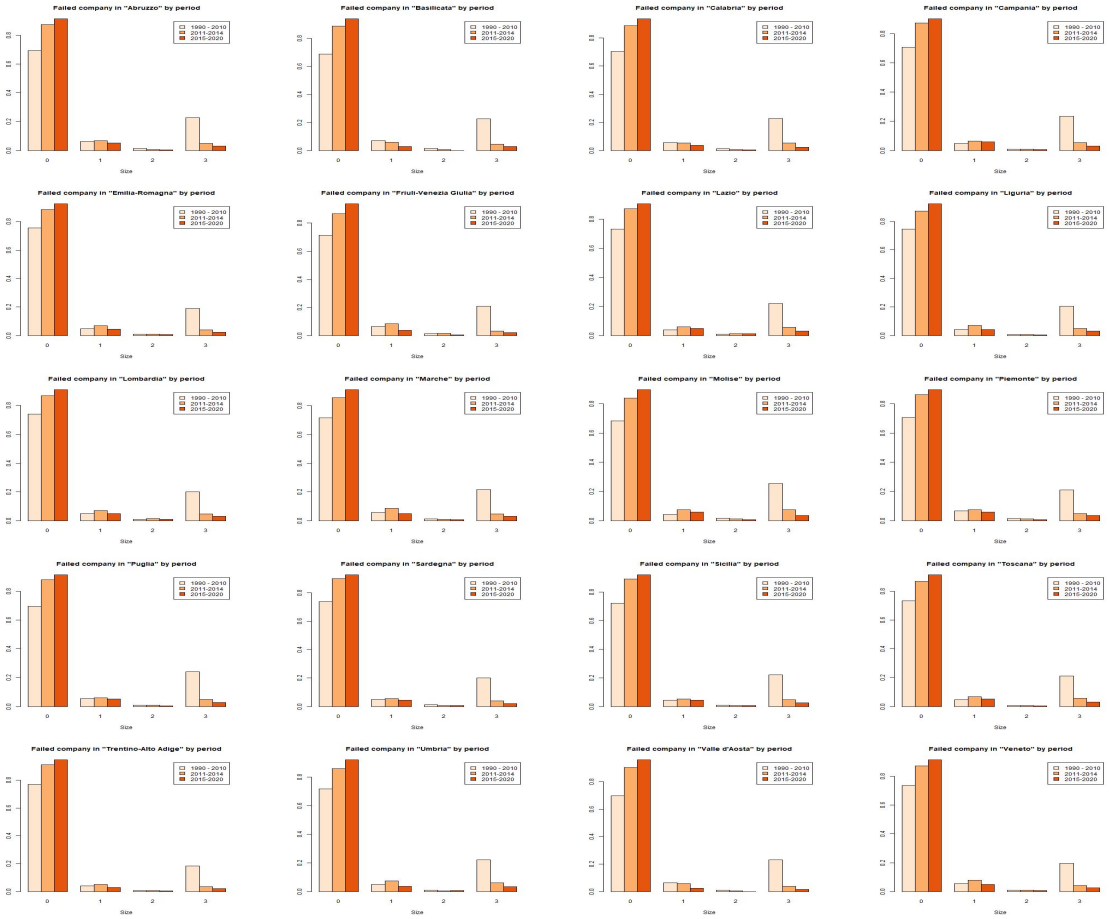


Figure 6.8: Size of Failed companies by year and region

6.3 Appendix B

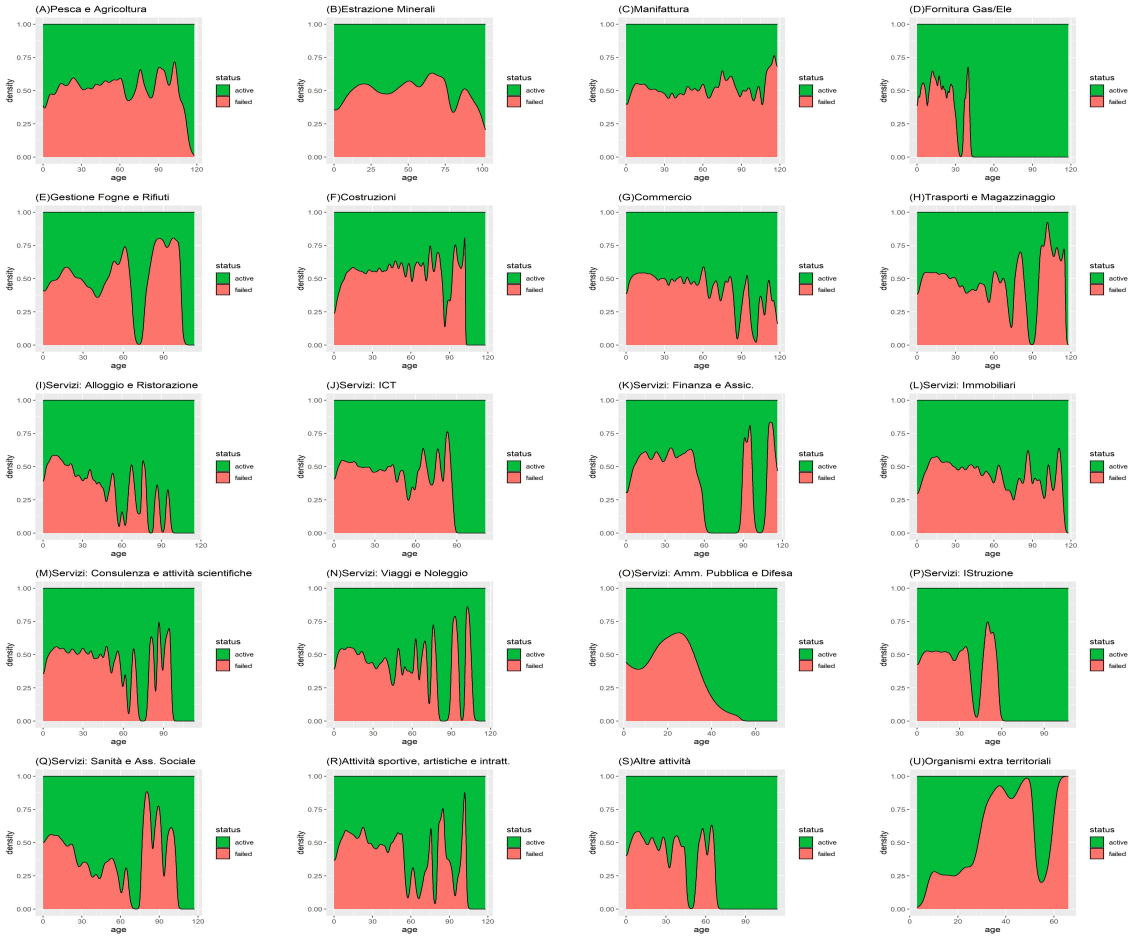


Figure 6.9: Age of Failed Companies by ateco sectors

Figure 6.10: Size of Failed Companies by ATECO sectors

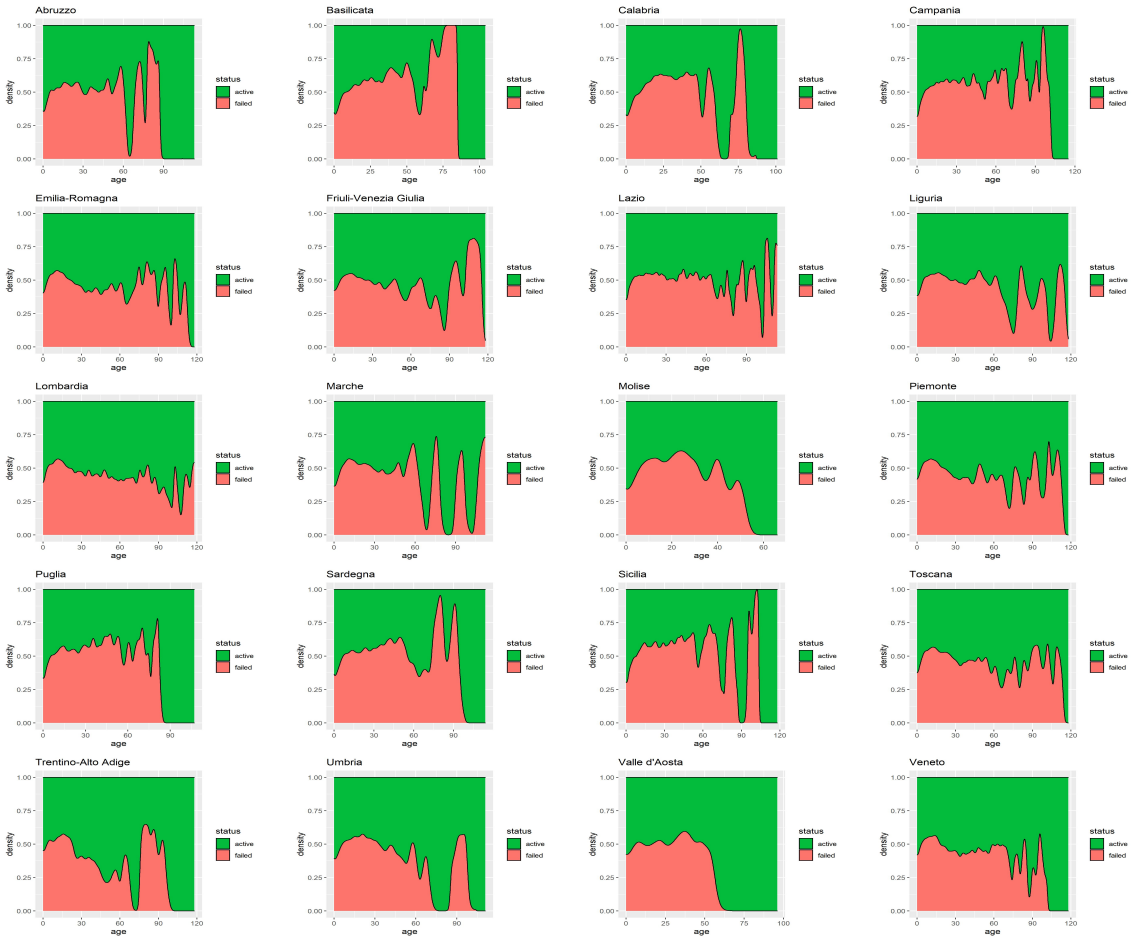


Figure 6.11: Age of Failed Companies by region

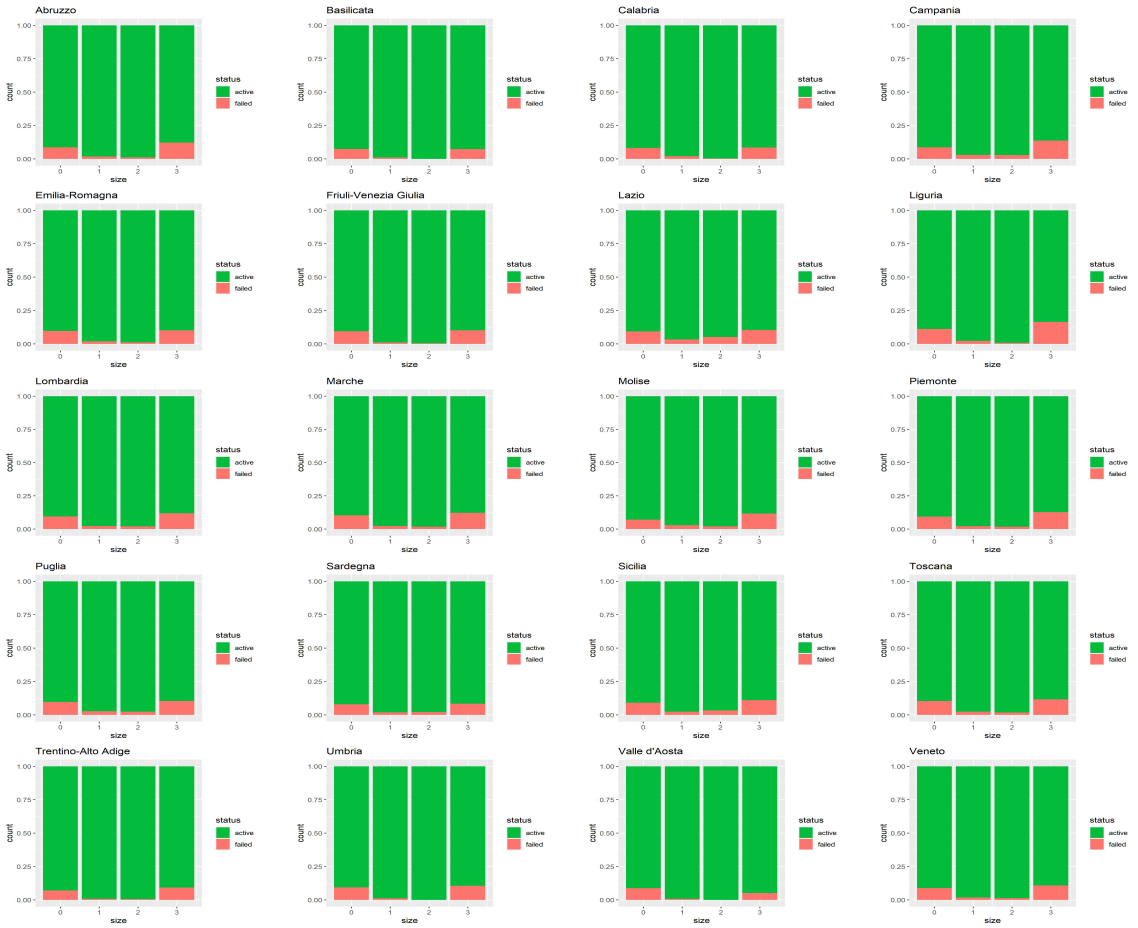


Figure 6.12: Size of Failed Companies by region