UNIVERSITÀ DI PISA

Statistics for Data Science
Project for A.Y. 2021/2022

RISK OF BUSINESS FAILURE
AIDA Dataset

Jordanos Feyissa Gemechu
Bruno Javier Limón Ávila
Carla Lorena Trejo Silva

Academic Year 2021-2022

Table of contents

1 Data Understanding & Preparation **2**
1.2 Missing values & outliers 5

2 Question A **6**
2.1 Age distribution analysis between failed and active companies 6
2.1.1 For specific company forms 6
2.1.2 For specific ATECO sectors 7
2.2 Size distribution analysis between failed and active companies 8
2.2.1 For specific company legal forms 8
2.2.2 For specific ATECO sectors 9
2.3 Liquidity distribution analysis between failed and active companies 10
2.3.1 For specific company forms 10
2.3.2 For specific ATECO sectors 11

3 Question B **12**
3.1 Age distribution analysis of failed companies over different years 12
3.1.1 For specific company forms 13
3.1.2 For specific locations 14
3.2 Size distribution analysis of failed companies over different years 15
3.2.1 For specific company forms 16
3.2.2 For specific locations 17

4 Question C **18**
4.1 Probability of failure conditional to Age 18
4.1.1 For specific company forms, ATECO sectors and locations 19
4.2 Probability of failure conditional to Size 19
4.2.1 For specific company forms, ATECO Sectors and locations 20
4.3 Probability of failure conditional to Liquidity 20
4.3.1 For specific company forms, ATECO Sectors and locations 20

5 Question D **21**
5.1 Developing a scoring and rating logistic regression model 21
5.1.1 Solving multicollinearity 22
5.1.2 Solving class imbalance 23
5.1.3 Model evaluation 23
5.1.4 Rating model alternative 25

6 Question E **26**
6.1 Confidence intervals of predictions in a logistic regression model 26

7 Conclusions **27**

References **28**

1 Data Understanding & Preparation

The following work has as general purpose to identify and analyze the factors that bring a company to failure, to do this we have analyzed the AIDA dataset, a compendium of historical financial indicators from many Italian companies along with their main characteristics, such as legal form or region. Specifically, the AIDA dataset is composed of 1,894,412 records with 80 variables, of which 71 are of numeric type, 6 factor and 3 character.

First, we took a look at the 6 factor variables to get a deeper understanding of the composition of these companies and in hopes of determining a definition of failure for our purposes. The business status refers to the situation of a company, its capacity to incur in business activities, and the possible state of insolvency it might be in. Within this dataset the possible legal statuses a company might have are: active, active (default of payments), active (receivership), bankruptcy, in liquidation, or dissolved.(See fig. 1 for their frequency distribution)

While some models for the prediction of business failure risk only used bankruptcy as a definition of failure. Balcaen and Ooghe point out that bankruptcy information might be contaminated by firms that do not show any real signs of failure and do not regard other causes for the failure of a business. Therefore, it was decided that in order to simplify the definition of failure, failure would be defined as any company that does not have an active status. With this definition, we created a new variable *Status*, enclosing every company in either an Active or Failed status, this can be seen in fig. 2.

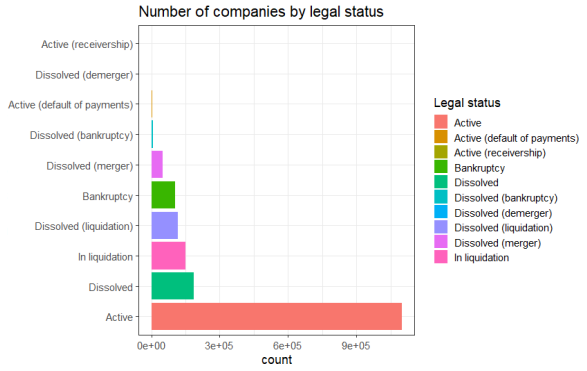


Figure 1 - Frequency of companies by legal status

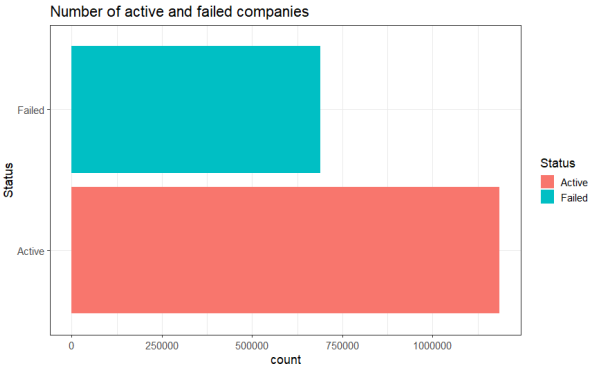


Figure 2 - Frequency of active and failed companies

Now having the active/failed status of every company, we can keep looking at the factor variables that can help us identify the particular type of company, for example, in fig. 3 we can appreciate the frequency of companies by region, along with fig. 4 for the frequency by ATECO sector code, which is a code related to the type of activity that a company carries out. Both of these figures have been splitted by the newly produced variable, *Status*, to see if there is a significant difference among these companies

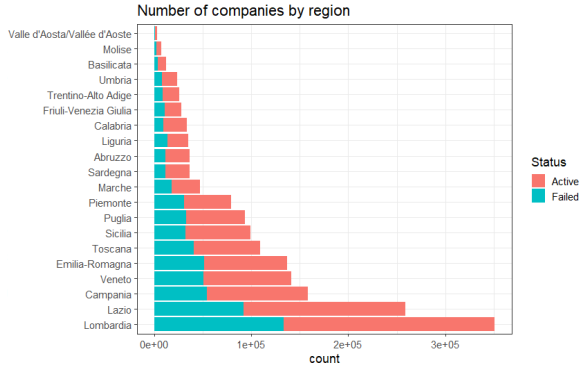


Figure 3 - Frequency of companies by region

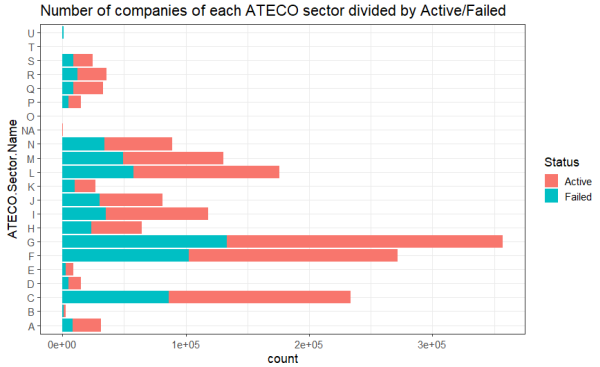


Figure 4 - Frequency of companies by ATECO sector

It is worth noting how in both of the previous figures there are elements with a very low frequency, such as ATECO code T or region Valle d'Aosta, along with elements with an undoubtedly bigger distribution within the dataset, such as region Lombardia or sector G. This differences might be of interest for further sections in which we will analyze how the distribution of a particular variable changes for specific types of companies.

Another interesting categorical variable is legal form, which similarly to the previous figures, shows a heavy inclination to some types of forms such as S.R.L., while having very low frequencies in the rest. (See fig. 5)

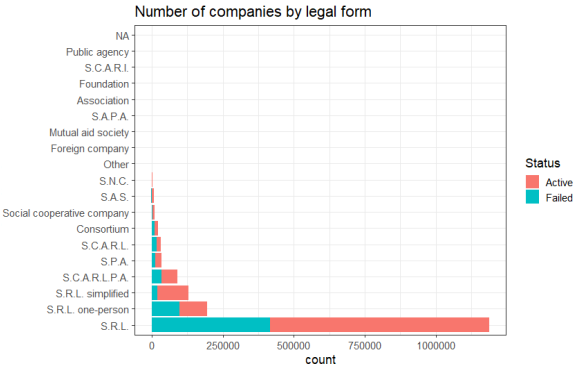


Figure 5 - Frequency of companies by legal form

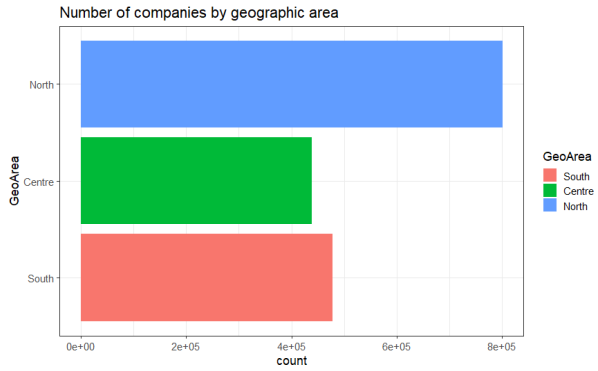
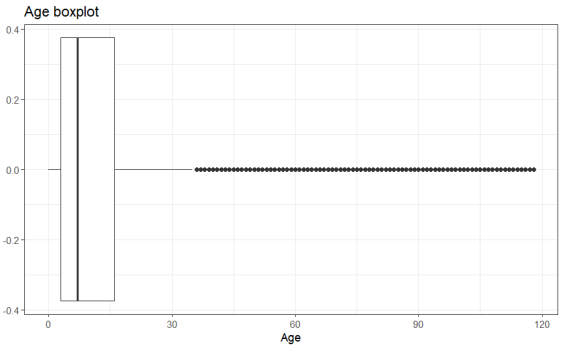


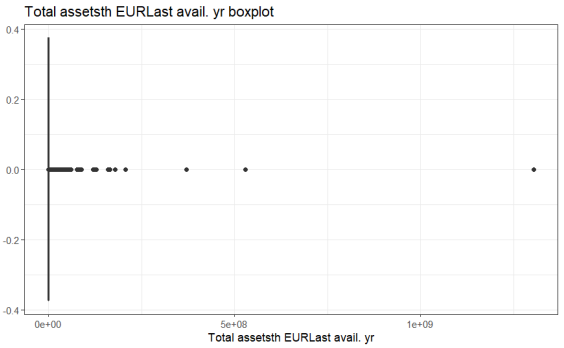
Figure 6 - Frequency of companies by geographic area

Lastly, in a similar fashion as what we did with legal status, we also put together companies belonging to a certain geographical location inside Italy, that is, north, center and south. Their frequencies can be seen in fig. 6, where we can see how the north hosts a big number of companies.

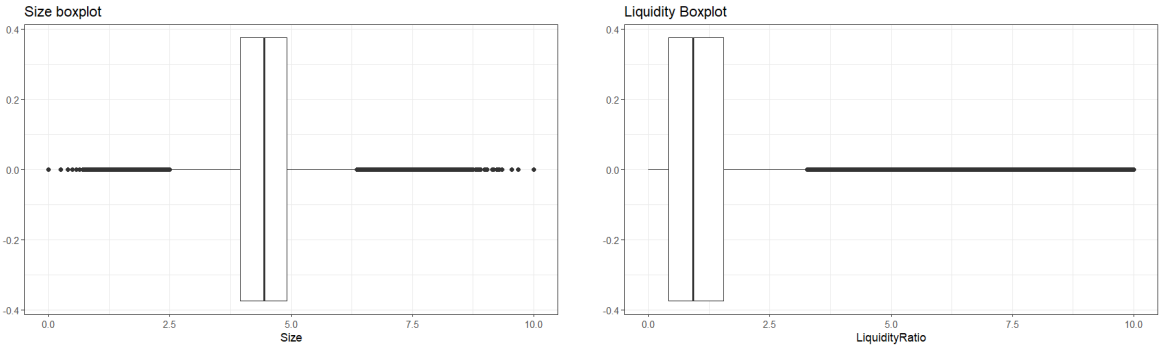
Moving on to the numerical variables, here we can find mostly financial indicators within a 3 year period, one different variable for each year, for example the first variable, *Banks/turnover%Lastavail.Yr* refers to the bank turnover for the last available year available for a particular company, then we have *Banks/turnover%Year - 1* and *Banks/turnover%Year - 2* for the 2 previous years, providing historical data. From these variables, there are also 2 that are not historical, mainly *Incorporation year* and *Last accounting closing date*, from which we can derive a new variable, *Age*, by subtracting the year of incorporation to the last accounting date. Then we also transformed the variable *Total assetsth EURLast avail. Yr* in order to get a representation of the size of a company, however, this variable, like many others within the dataset are heavily skewed towards very big values, we can observe this behavior for example by plotting its boxplot, where there is practically no interquartile range, and must of the data points are condensed within the first segment (seen in fig. 7B). This kind of skewness would affect any analysis done on the distribution of the variable, thus we decided to instead create a new variable, *Size*, considering its logarithm plus a constant to avoid values equal to 0 to become *-inf*, then we normalized this result into a scale from 0 to 10, for greater interpretability, the results are more much favorable, as seen in fig. 7C. Finally, this same normalization was applied to *Liquidity ratioLast avail. Yr* to get a more clear value of liquidity to use for further analyses. Table 1 summarizes these Boxplots.



A) Age Boxplot



B) Total assetsth EURLast avail. YrBoxplot



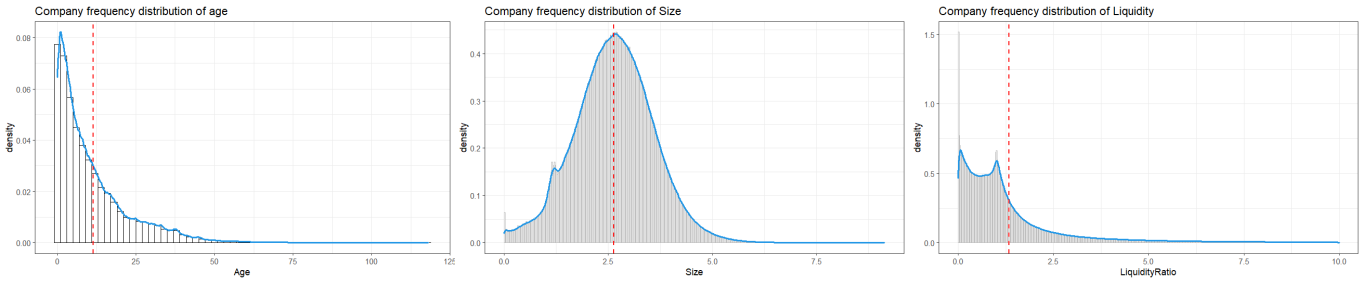
C) Size Boxplot *D) LiquidityRatio Boxplot*

Figure 7 - Boxplot of different variables to use for further analyses

	Age	Size	LiquidityRatio
Min	0	0	0
Max	118	9.12	10
Mean	11.41	2.63	1.34
Q1	1.00	1.40	0.13
Q3	28.00	3.81	3.04

Table 1 - Boxplot comparison for engineered features

In addition to Boxplots, we can also get valuable insights with the use of histograms, more or less confirming what we have previously seen, in fig. 8, we can see the the frequency distribution, the probability density function (blue line) and the mean (red line) of the previously analyzed variables.

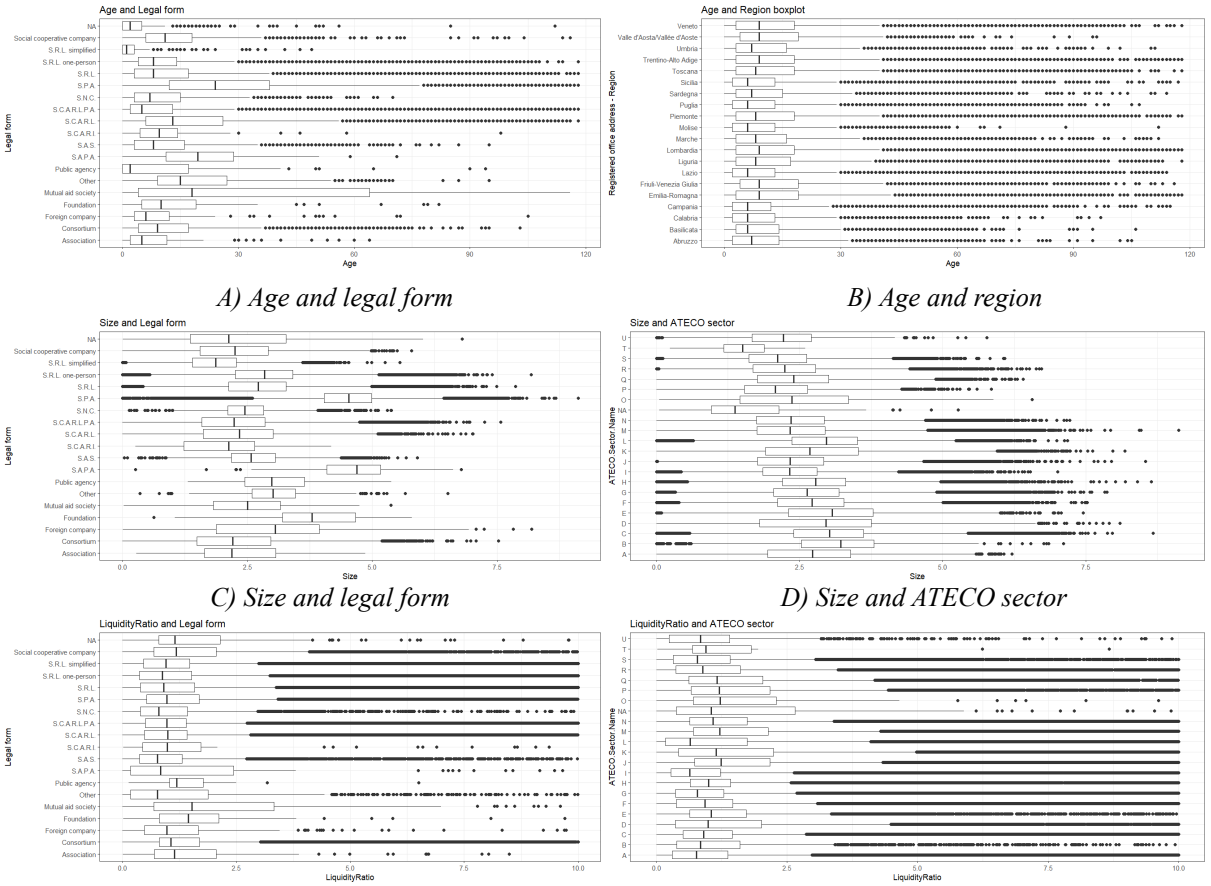


A) Age distribution *B) Size distribution* *C) LiquidityRatio distribution*

Figure 8 - Histogram of analyzed variables, along with density function and mean

From both fig. 8A and 8C we can confirm that both *Age* and *LiquidityRatio* seem to have a positive skewness, while *Size* seems almost symmetrical, except for a heavier tail on the right side. However, to not only rely on visual analysis, we also decided to calculate the skewness and kurtosis values with the help of the functions *skewness* and *kurtosis*, both from the library *moments*. These functions gave us the following results, for variable *Age*, skew = 2.01, kurt = 9.05, for *Size*, skew = 0.05, kurt = 3.27 and for *LiquidityRatio*, skew = 2.61, kurt = 11.01. These results confirm the visual analysis, showing that both *Age* and *LiquidityRatio* have positive skewness or are right-skewed and also have positive kurtosis, i.e. they are both leptokurtic, whereas *Size*, with a skewness so close to 0, means that it has almost no skew, if only ever so slightly right-skewed. As for kurtosis, with a value close to 3, it is mesokurtic, confirming it belongs to a normal distribution.

Finally, these variables were confronted to some of the categorical ones with the use of boxplots, to see if there is any interesting behavior in their relationship that could be useful for future analysis. Some of these results can be seen in fig. 9.



A) Age and legal form *B) Age and region*

C) Size and legal form *D) Size and ATECO sector*

E) LiquidityRatio and legal form *E) LiquidityRatio and ATECO sector*

Figure 9 - Relationship between different numerical and categorical attributes

By looking at these figures we can understand how these variables interact with each other, for example in the case of legal form, we can see that S.P.A. companies tend to be older and bigger than the rest of them. These kinds of insights might be valuable for further sections of this work.

1.2 Missing values & outliers

Regarding missing values, with the help of the function *skim* from library *skimr*, we can get a summary of every variable with some of its main characteristics, one of these is n_missing, telling us the amount of *NA* for each variable. This resulted in some worrying results, since most of the numeric values suffer from substantial amounts of missing values, sometimes even reaching more than 75% of *NA*, such as in the case of Cost of debit (%)%Last avail. Yr. Upon further inspection, we can see that these missing values occur not at random, and some variables have them way more than others, meaning that there is a relationship between variables and their missing values, this could result in bias, depending on the type of analysis/model performed. A solution to this is to handle the missing values, either completely removing them, which can affect the validity of analysis/model due to low amounts of data available, or replacing them with a value such as the mean, however, this approach however can be negatively influenced by the presence of outliers, which can affect values of central tendency such as the mean, as an alternative, the mode or median could be used. As for outliers, having a quick visual inspection of the boxplots previously reported, as in *Age* for example, we can see that a big portion of the distribution is considered as an outlier, due to most of the companies having less than 30 years of age. This however does not mean we should get rid of them, since that would remove an important portion of the data that could be useful. Instead, we shall deal with them accordingly according to the needs of the next methods proposed.

2 Question A

After getting to know the dataset, it is time to start the evaluation of the proposed questions. We begin by comparing the distributions of age, size, and liquidity between failed and active companies at a specific year. In this case the year that will be used is 2018, as it is the year with the largest amount of records.

2.1 Age distribution analysis between failed and active companies

The first of the attributes to be evaluated is the age of the active companies in 2018. In this case, we can compare the ages of the active and failed companies in 2018.

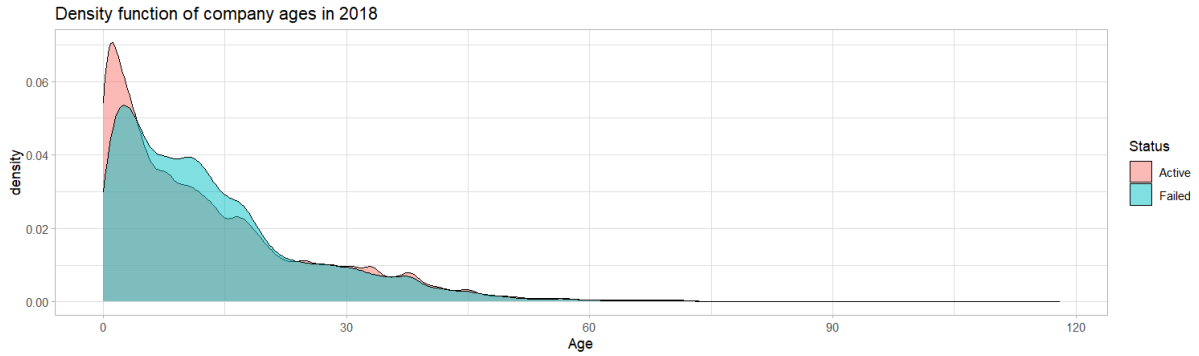


Figure 10 - Distribution of company ages in 2018

In this analysis of age distribution it is possible to observe that active companies are newer or have been working from recent years. The more we move to the left, the higher is the density of active companies. As for the failed companies it has a similar distribution, where there is a higher density of failed companies in the younger side, and as they get older there is a smaller density.

2.1.1 For specific company forms

To compare the age distribution of active and failed companies in a given year, we decided to use the t-test.

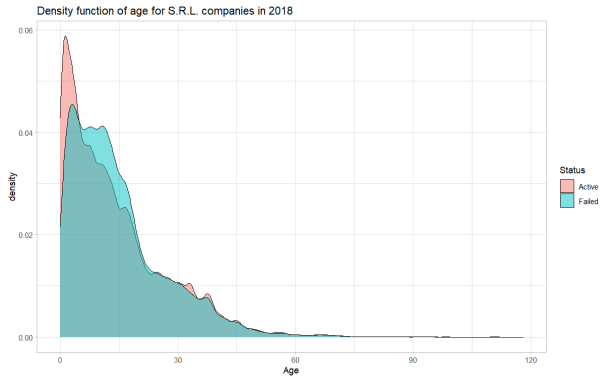


Figure 11 - Density function of age for S.R.L.

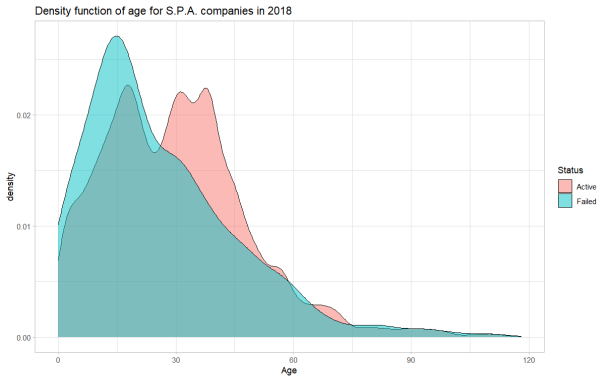


Figure 12 - Density function of age for S.P.A.

As the number of observations is really large, we could use the t-test even though the distribution does not follow a normal distribution. It was not possible to use the Kolmogorov-Smirnov test because Age distribution is not continuous. For the t-test, the value of $\alpha = 0{:}05$, rejecting the null hypothesis if and only if the p-value is less than α.

Legal form	Test	Hypothesis	p-value
S.R.L.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	2.2e-16
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1
S.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	6.128e-09

Table 2 - μ_1 is the real average of the age of the failed companies in 2017.
μ_2 is the real average of the age of the failed companies in 2018.

From an overall analysis of age distribution among active and failed companies we pass to a deeper analysis to see the difference of age distribution between a Limited liability company (S.R.L) and a Public limited company (S.P.A).
In the Limited liability company (S.R.L) as we have seen earlier, active companies are newer and have a higher percentage of density with respect to old ones; but for the Public limited company (S.P.A) is different. Both active and failed companies have almost the same trend at starting point in time, then there is an increase in active companies that will be covered after by failed companies.

2.1.2 For specific ATECO sectors

Equivalently to the previous point, now the data is divided by the ATECO sector code it belongs to. Fig. 13 and 14.

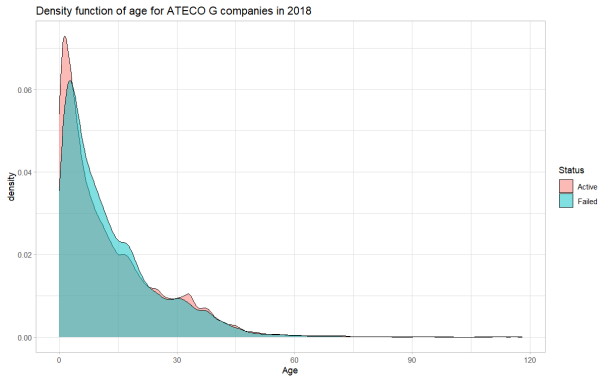


Figure 13 - Density function of age for ATECO G

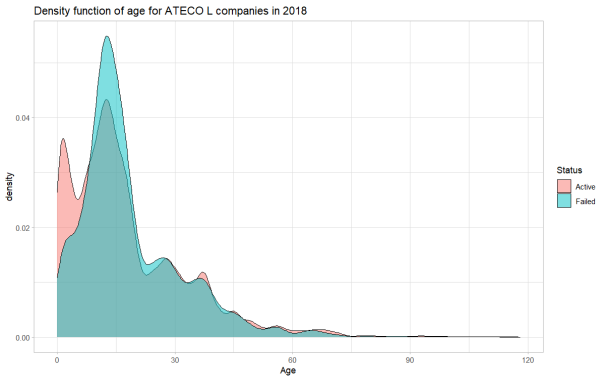


Figure 14 - Density function of age for ATECO L

The difference of age distribution between sectors is visible. With the first one we can see that there is an increasing growth of both active and failed companies, where the active one in the recent year has a higher density. While in the (second sector) like the first one have the same growing trend but at some point, the density of failed companies tends to increase and then to decrease, allowing the active companies to have a higher density with respect to it.
This can be further observed by the results in testing, where the alternative hypothesis is accepted under less diverse results than in tests for other sectors, such as L, or attributes such as company legal form. The results are shown in table 3.

ATECO	Test	Hypothesis	p-value
ATECO G	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.3242
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.6758
ATECO L	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	5.746e-05
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9999

Table 3 - Results of t-test for different ATECO sectors and age of companies in 2018

2.2 Size distribution analysis between failed and active companies

Similar to the age distribution, the next step in the analysis is the testing of the distribution of companies in function of the size.

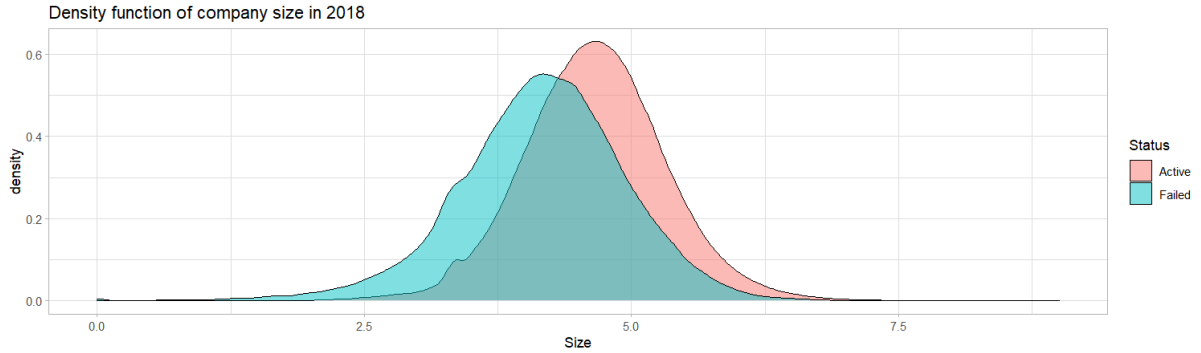


Figure 15 - Density function of company size in 2018

Here we can observe the size distribution of active and failed companies. They have the same shape, although the number of active companies is higher compared to the failed one.

2.2.1 For specific company legal forms

Similarly as previously done, the analysis continues with the size distribution for the legal forms of these companies. Figure 16 and 17 show the distribution for these records.

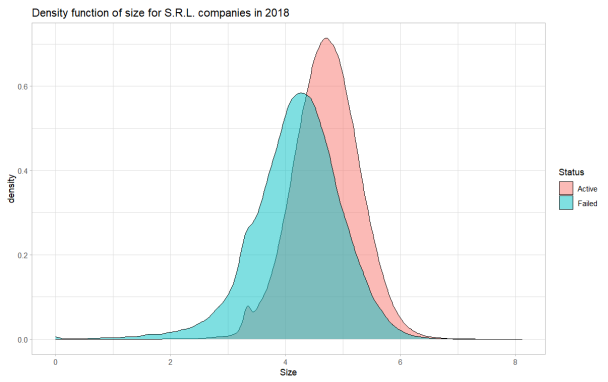


Figure 16 - Distribution for S.R.L companies sizes

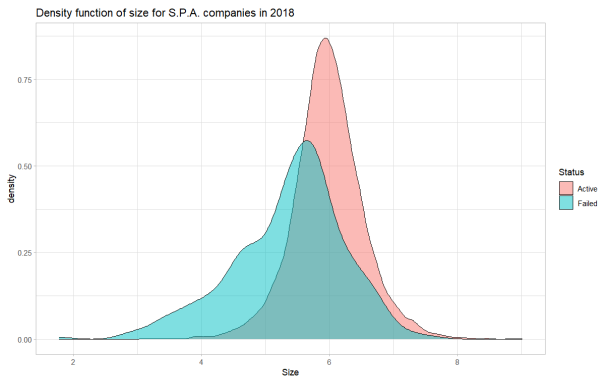


Figure 17 - Distribution for S.P.A. companies size

The size distribution of active and failed companies for Limited Liability company (S.R.L) and Public limited company (S.P.A.) differs. In the Limited Liability company (S.R.L) the gap between active and failed companies is reduced; even though the failed one has a small size and it is less dense with respect to the active one.
In the Public limited company (S.P.A.), the size of the active companies is larger and more densely distributed compared to the failed ones. When the size of failed companies starts to decrease the size of active companies continues to grow and arrives at a certain level and starts to decrease.
The results of the evaluation by t-test of active and failed companies in 2018 are consistent with this. In table 2 are presented the results for the two most recurrent legal forms in this year. However, the results for the rest can be seen in the code. If a legal form had under 100 observations, it was arbitrarily not evaluated as it had two few data.

Legal form	Test	Hypothesis	p-value
S.R.L.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_2 \leq \mu_1 \leq \mu_2$	2.2e-16
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	2.2e-16
S.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_2 \leq \mu_1 \leq \mu_2$	2.2e-16
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	2.2e-16

Table 4 - Results of t-test for different legal forms of companies in 2018

2.2.2 For specific ATECO sectors

As the previous point, the same continues for the size distribution of companies for the different ATECO sectors. In this case there is a focus on the sectors G and L for this study, as represented by the following figures.

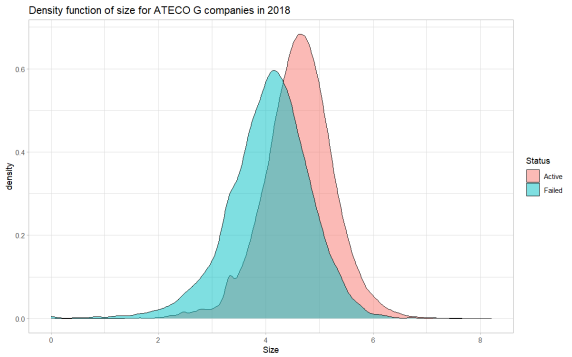


Figure 18 - Density for liquidity of ATEGO G Companies by size in 2018

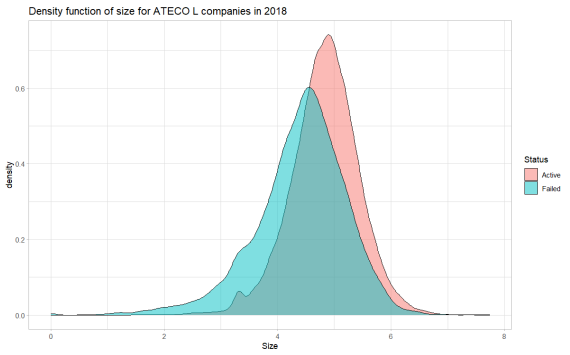


Figure 19 - Density for liquidity of ATEGO L Companies by size in 2018

While the distribution of size for specific sectors has almost the same growing trend, the active one prevails.

ATECO	Test	Hypothesis	p-value
ATECO G	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	2.2e-16
ATECO L	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.5
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.5

Table 5 - Results of t-test for different ATECO sectors and size of companies in 2018

In this case, we can observe that for sector G the tests performed as predicted, however, when we get to sector L we observe that in both our tests the p-value was the same value as the value of α. In this case further tests could be implemented to get a more satisfying answer.

In these teste we can observe that for some cases while the condition …

2.3 Liquidity distribution analysis between failed and active companies

The next distribution to be evaluated is the one of liquidity ratio, as defined in the previous section of the analysis.

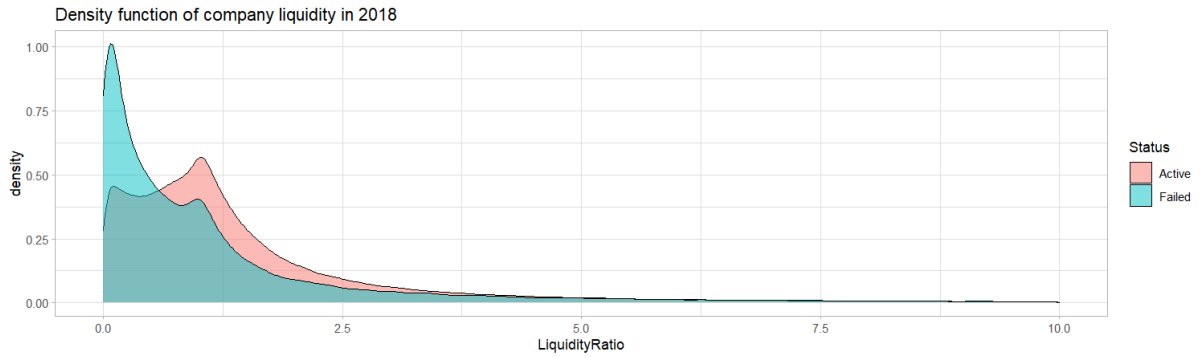


Figure 20 - Density function of company liquidity in 2018.

Liquidity ratios are an important class of financial metrics used to determine a debtor's ability to pay off current debt obligations without raising external capital.
The value of liquidity ratio of most failed companies is in the range of 0 and 2.5, from this we can infer that most of them are unable to cover short term obligations, since the value between 0 and 1 is denser. While the liquidity ratio for the active companies have the same trend, but it is less dense.

2.3.1 For specific company forms

The analysis according to the legal form and liquidity was further enhanced by use of a boxplot to see the distribution of the different legal forms across Liquidity as can be seen in fig. 9E.
The liquidity ratio for the Limited liability company for failed companies starts at higher density, then to decrease as the ratio value increases. Even Though the active companies have a lower density, most of the companies fall in the range of 0 and 2.5.

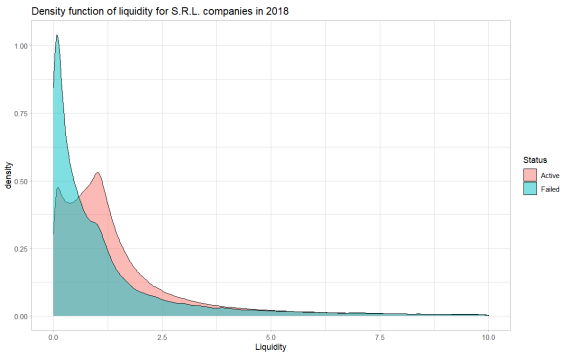


Figure 21 - Density for liquidity of S.R.L. Companies by size in 2018

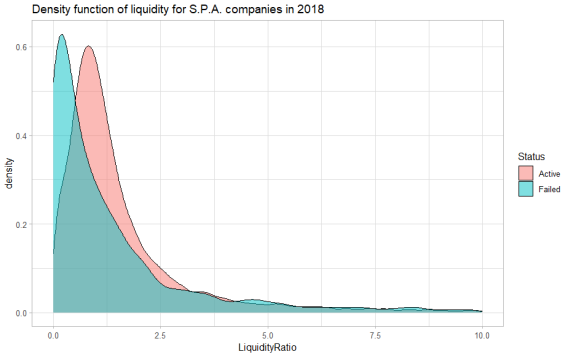


Figure 22 - Density for liquidity of S.P.A. Companies by size in 2018

While in the Limited partnership apart from having the same tendency; the liquidity ratio of failed companies does not have a continuous trend, where it stops at a certain point and then restarts at a higher ratio with lower density, then falling down.

Legal form	Test	Hypothesis	p-value
S.R.L.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	2.2e-16
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1
S.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	6.128e-09

Table 6 - Results of t-test for different legal forms and liquidity of companies in 2018

2.3.2 For specific ATECO sectors

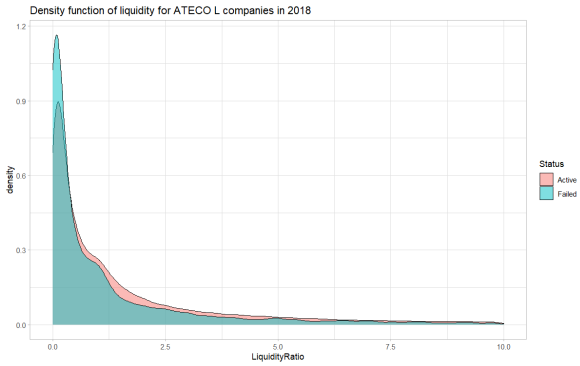


Figure 23 - Density for liquidity of ATEGO G Companies by size in 2018

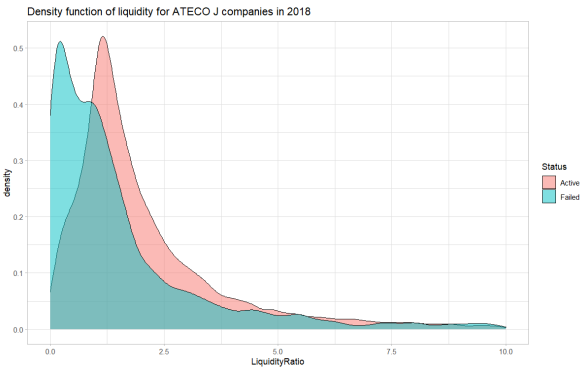


Figure 24 - Density for liquidity of ATEGO L Companies by size in 2018

The results for this set of tests are presented in table 7. The distribution of liquidity ratio for sector J of failed companies is denser at a lower value of the ratio, demonstrating the inability of the company to handle short term obligations.

ATECO	Test	Hypothesis	p-value
ATECO J	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.3853
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.6147
ATECO L	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	5.746e-05
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9999

Table 7 - Results of t-test for different ATECO sectors and liquidity of companies in 2018

While for the active companies it starts at a lower density and then increases as the value raises, where eventually it continues to decrease. While for the sector L is different as the ratio increases the density tends to decrease for both active and failed companies.

3 Question B

The next point is the analysis is expanding the analysis of the previous point from only one year to further years. In this case, the analysis was implemented for 2018, which we used in the previous point and the precedent year, 2017. Unlike the previous analysis point, here only the companies that have a failure status were considered.

3.1 Age distribution analysis of failed companies over different years

In order to compare the distribution of the ages of failed companies over this period of time. It was decided to use the t-test. Although the age does not follow a normal distribution, it is possible to apply the t-test as the number of observations is very large. Unfortunately, it is not possible to use the Kolmogorov-Smirnov test because the age at failure distribution is not continuous. For the tests, an arbitrary level of significance was selected in α = 0:05, rejecting the null hypothesis if and only if the p-value is less than α.

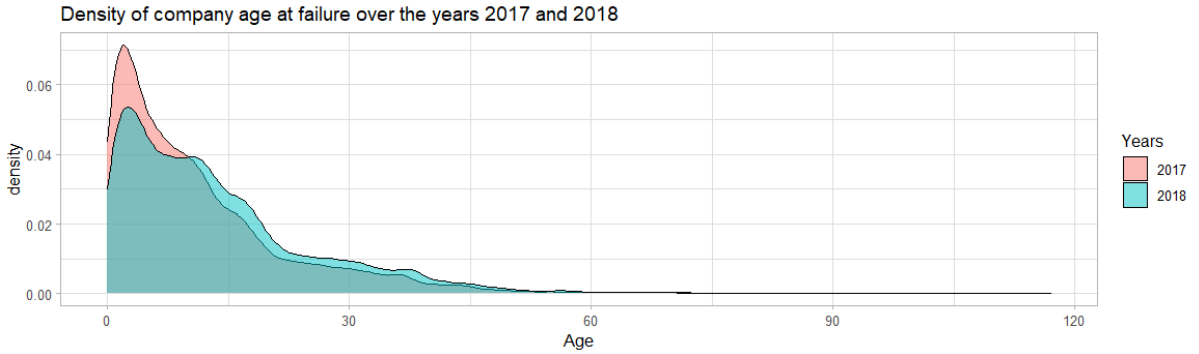


Figure 25 - Distribution of companies by age at failure during the selected years.

It was possible to observe that in both years, the distributions resulted similarly, with a higher incidence of failure in younger companies for both years..

Test	Hypothesis	p-value
t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	2.2e-16
t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1

Table 8 - 1 is the real average of the age of the failed companies in 2017. 2 is the real average of the age of the failed companies in 2018

Further proving the information presented by the graph, our null hypothesis is discarded in the first scenario, as the average age of the companies with a failure status in 2017 is smaller than that of 2018.

3.1.1 For specific company forms

Once the analysis of the distribution of company ages at failure in these two years was evaluated, it was required to evaluate the distribution of company ages at failure for the different legal forms a company might have in Italy.
In the AIDA dataset, there are listed over 15 possible legal forms of companies in Italy, figures 27 and 28 portray the distribution for these different forms and their ages for the years 2017 and 2018, respectively.

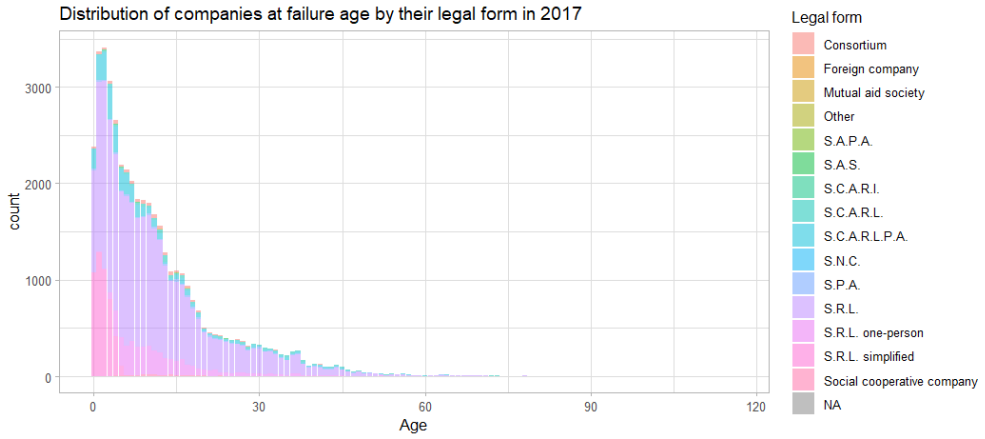


Figure 26- Distribution of companies at failure age by their legal form in 2017.

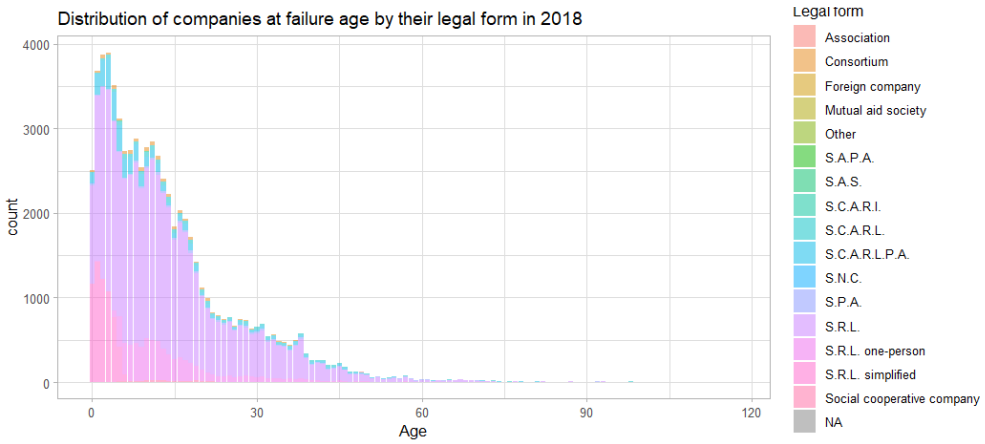


Figure 27 - Distribution of companies at failure age by their legal form in 2018.

However, as can be perceived in these figures, the distribution of by legal form is highly concentrated in several forms, therefore a further filter was used in order to limit the testing to the legal forms that had over 100 observations. With this information, it is possible to test the distribution for most of the forms.

Legal form	Test	Hypothesis	p-value
S.R.L.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	2.2e-16
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1
S.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.005645
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9944
S.C.A.R.L.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1.234e-15
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1

Table 9- μ_1 is the real average of the age of the failed companies in 2017. μ_2 is the real average of the age of the failed companies in 2018.

In these cases, we can see that it behaves similarly, as the general trend of all different legal forms are mostly similar to the one when we didn't divide by the different legal forms. Table 3 presents some of the results, further results can be found in the code. Amongst the legal forms that didn't get tested due to a lack of observations are foreign companies, S.A.P.A., and associations.

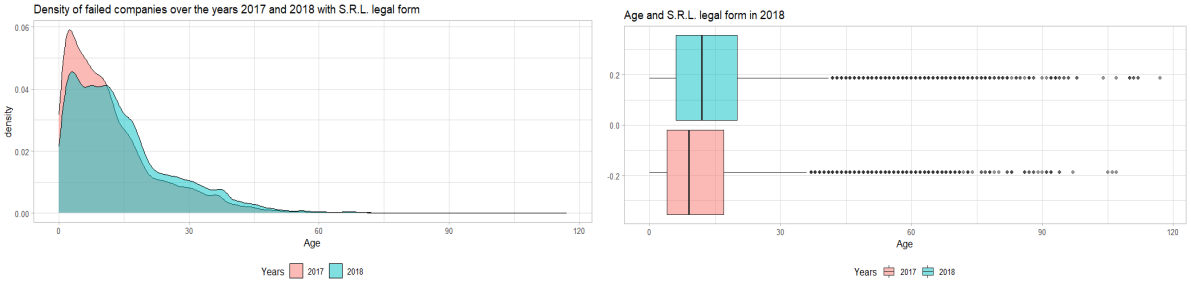


Figure 28 - Distribution for S.R.L. companies with failure status by age in 2017 and 2018

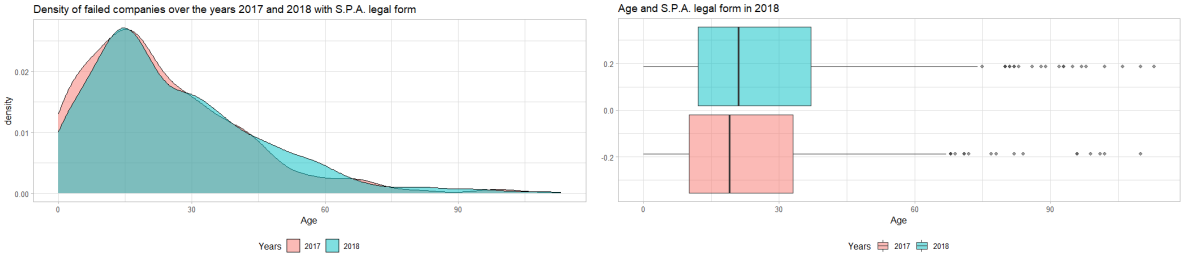


Figure 29 - Distribution for S.P.A. companies with failure status by age in 2017 and 2018

3.1.2 For specific locations

Similarly to the previous point, we can get the distribution for the failure age of companies divided by the region they were registered in. In this case, the division was done by considering the 20 regions Italy is divided. The first visualization (fig. 30) refers to the distribution of failed companies by region in 2017 and 2018.

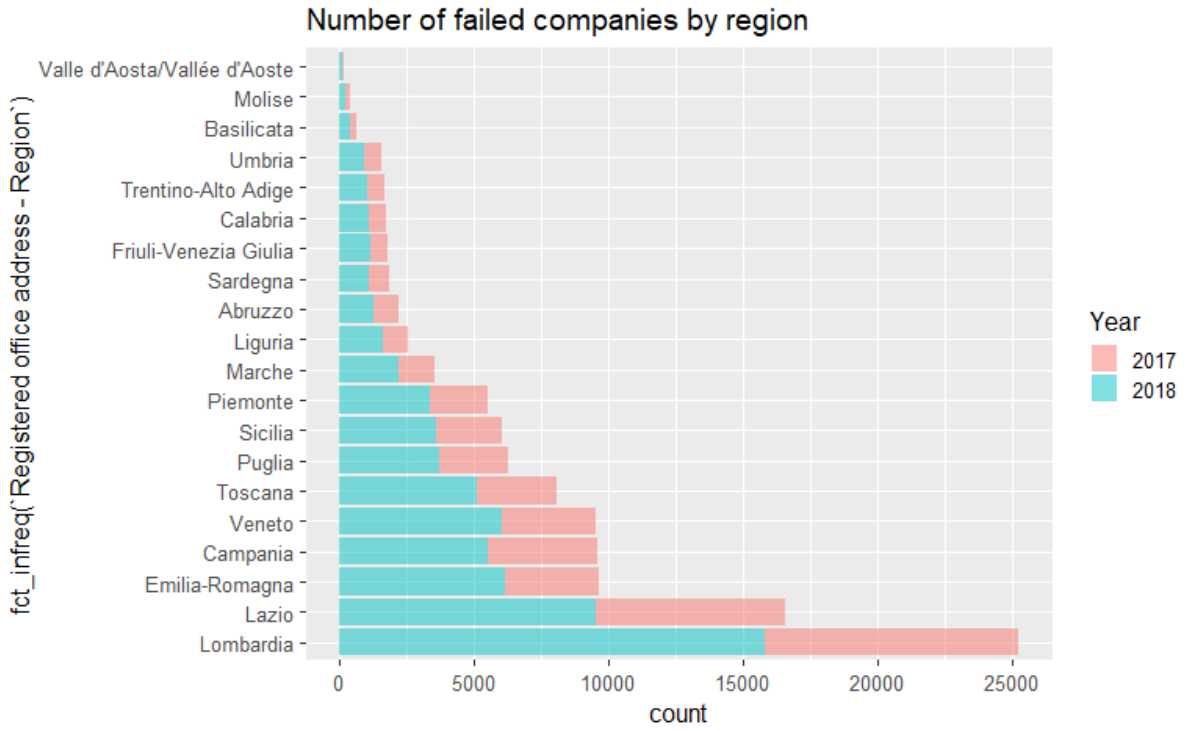


Figure 30 - Distribution of failed companies by region in 2017 and 2018.

3.2 Size distribution analysis of failed companies over different years

Unlike the previous case, the analysis of company size at failure of companies in this period of time, we get a normal distribution. In this case it was possible to also evaluate using the Kolmogorov-Smirnov (ks) test, in addition to the t-test of the previous section. In both cases, the value of α that against which the p.value will be evaluated, will always be α = 0.5.

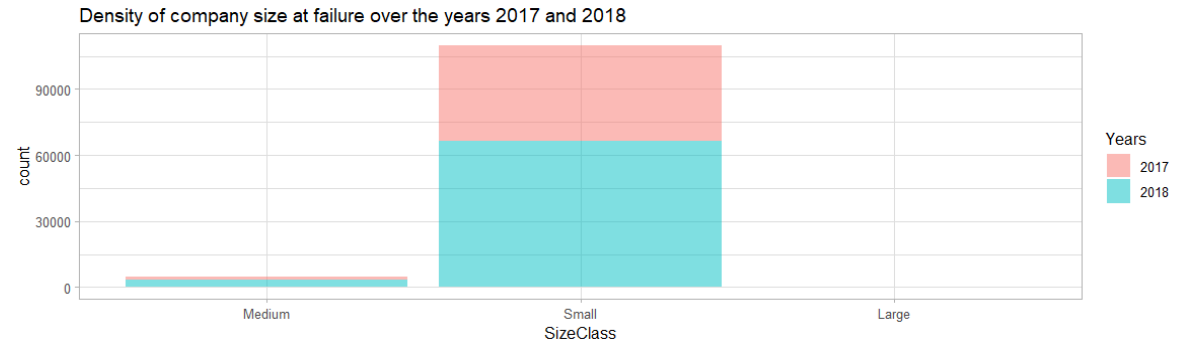


Figure 31 - Distribution of companies by company size class at failure during the selected years

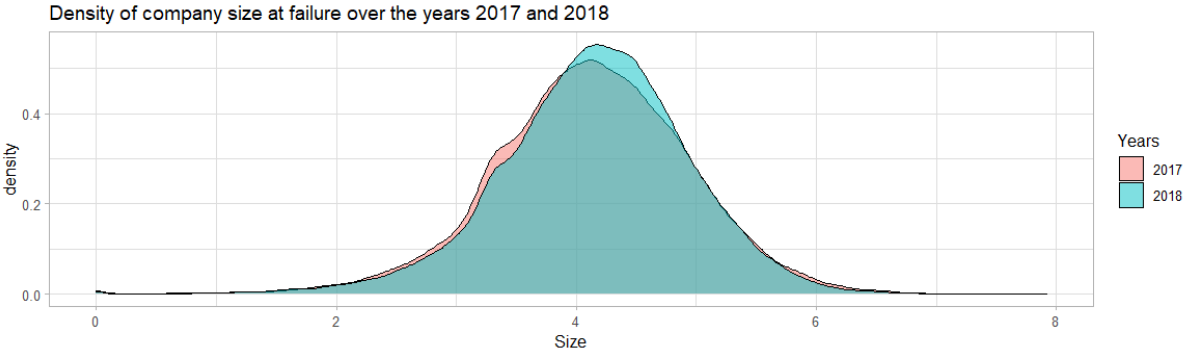


Figure 32 - Distribution of companies by company size at failure during the selected years

Test	Hypothesis	p-value
t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_2 \leq \mu_1 \leq \mu_2$	1.781e-08
t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	8.903e-09
t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1
ks-test	H_0: *samples come from the same distribution* H_1: *samples come from different distributions*	0.07254

Table 10 - μ_1 is the real average of the age of the failed companies in 2017. μ_2 is the real average of the age of the failed companies in 2018

The t-test in However, the results of the t-test and the ks- test are clashing. As the ks-test returns a value p-value < α. On the other hand, all the three scenarios always return a p-value greater than α. This allows us to discard the null hypothesis and to affirm that the averages of the distributions are on average equal.
One possible cause of these conflicting results could be the way the tests compare distributions; the t-test uses the population mean, while the Kolmogorov-Smirnov is based on the distance between the empirical distribution functions of the two samples.

3.2.1 For specific company forms

Similarly as done before, now we are evaluating for the selected years only the failed companies and their legal form. While the results of the different legal forms can be seen in the code, as long as they have more than 100 records each, in this report we are only presenting the results for the S.R.L. and S.P.A. failed companies.

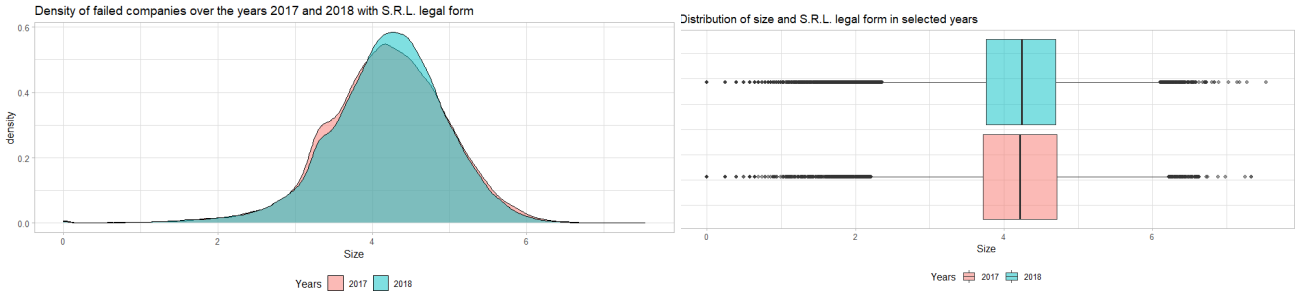


Figure 33 - Distribution for S.R.L. companies with failure status by size in 2017 and 2018

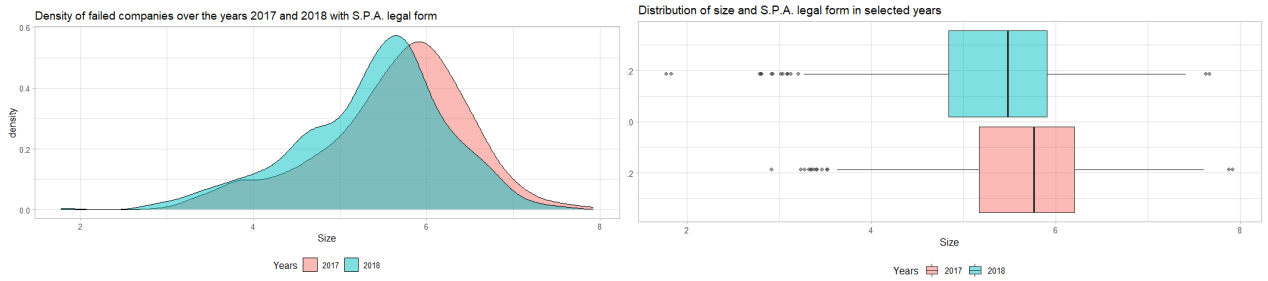


Figure 34 - Distribution for S.P.A. companies with failure status by size in 2017 and 2018

Among the most significant interesting results we have that for the S.R.L.type, the p-value returned by the test allows us to discard the null hypothesis in favor of the alternative hypothesis left unilateral. In the other hand, for the S.P.A form, the solution is not as clearly given using the t-test.

Legal form	Test	Hypothesis	p-value
S.R.L.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.07724
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9228
S.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	1
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	1.647e-08
S.C.A.R.L.P.A.	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.0001597
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9998

Table 11 - μ_1 is the real average of the age of the failed companies in 2017. μ_2 is the real average of the age of the failed companies in 2018

3.2.2 For specific locations

As for the distribution of specific years for companies of different sizes for the various regions of Italy, we get a distribution where the vast majority of companies belong to the small category, as seen in the fig. 35 and 36.

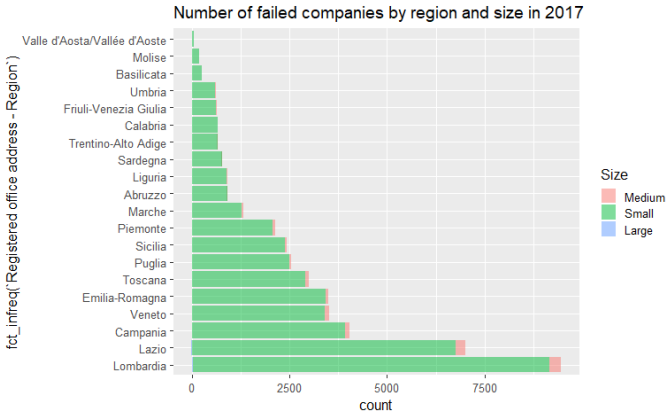


Figure 35 Distribution of failed companies by region and size in 2017

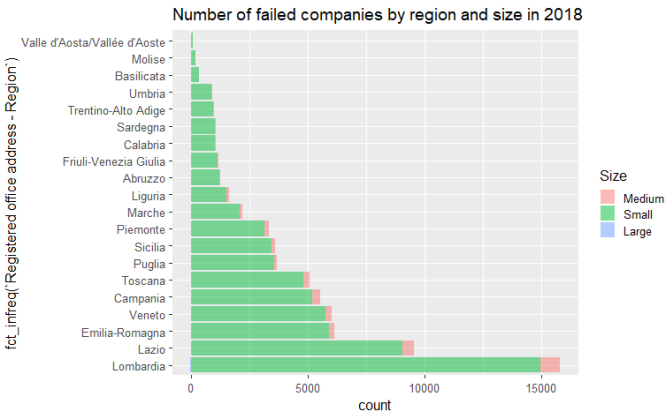


Figure 36 - Distribution of failed companies by region and size in 2018

As the Regions vary, all the tests carried out do not permit the elimination of all null hypotheses in favor of the alternatives. However, looking at the value of the p-value one can still understand what the relationship is, even if not significant, for the distributions considered. As can be shown in Table 6.

Region	Test	Hypothesis	p-value
Lombardia	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.03173
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9683
Lazio	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.3964
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.6036
Emilia-Romagna	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.06273
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.9373
Toscana	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	0.5715
	t-test	$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	0.4285

Table 12 - μ_1 is the real average of the age of the failed companies by region in 2017. μ_2 is the real average of the age of the failed companies by region in 2018

4 Question C

For this section we shall analyze the probability of failure at a given year conditional to the variables *Age*, *Size* and *LiquidityRatio* and see if they change for specific company forms, ATECO sectors or locations. This probability is known as conditional probability and takes into account is an event has occurred or is a variable has a certain value, in our case, we are looking at P(Status = Failed|Var = x).

4.1 Probability of failure conditional to Age

To compute this, we generated a table containing the frequency of failed and active companies for every age during 2017, then obtained the ratio of failed to total companies for each value, resulting in a final table containing the age, number of failed companies, number of active companies and probability of failure for that particular age. By plotting this probability along all possible values of x we get the following figure.

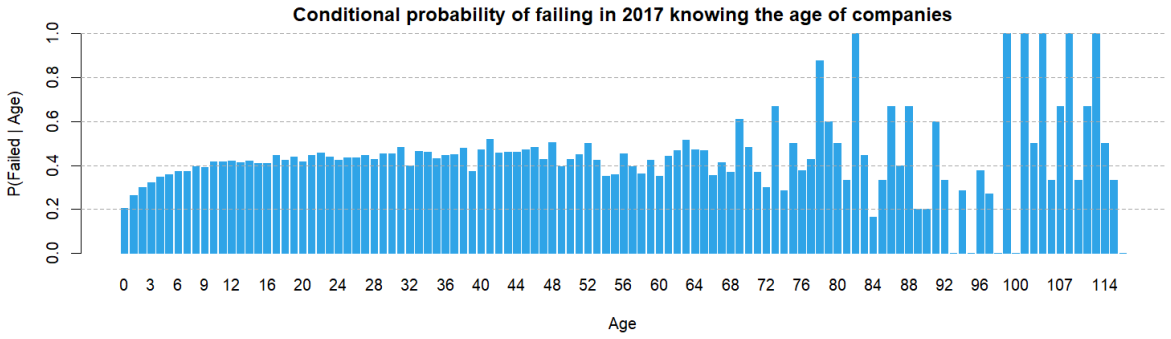
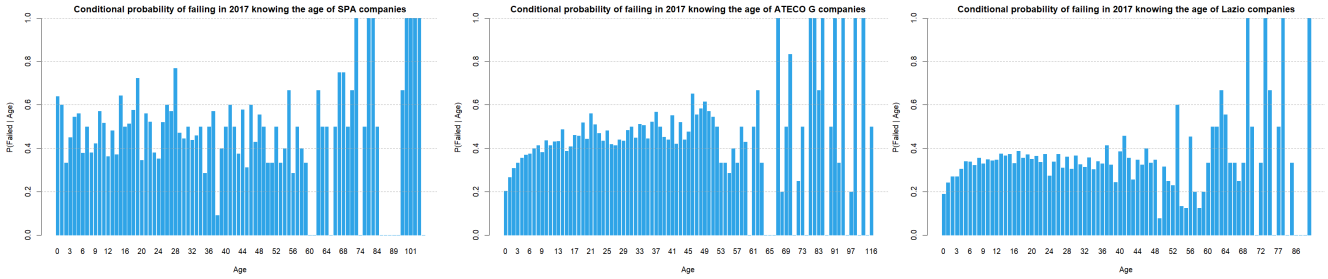


Figure 37 - Probability of failing in 2017 knowing the age of the companies

From visual inspection, we can see there's a slightly increasing trend upwards during ages 0-10, then stabilizing for the most part, except when reaching about the 7o age mark, where the probabilities start to go way up or down. Upon further inspection of the table of probabilities, we can see that this is due to the fact that the amount of available companies for certain ages are very low, going from the thousands at first to reaching only about a dozen later on, having cases where there's even only one company, either active or failed, which of course will result in 0 or 1 probabilities. For this reason this old-age probabilities might not be especially trustworthy because of their lack of data.

4.1.1 For specific company forms, ATECO sectors and locations

Moving on to further filtering of his analysis, we repeated the process but now adding additional variables, legal form, ATECO sector and region. For example analyzing only S.R.L companies or Lazio-based ones. The exact choice of these types of companies was in part aided by the initial analysis of variables, as in fig. 9, selecting for example S.R.L companies which share more or less the average age of most company forms and S.P.A., which have older companies in general. The results however don't seem to change significantly, and in some cases such as S.P.A. companies seem inconsistent again due to lack of data. Some of these results can be seen in fig. 38.



A) S.P.A. *B) ATECO G* *C) Lazio*

Figure 38 - Conditional probability of failing in 2017 knowing the age of the company and another attribute

4.2 Probability of failure conditional to Size

For the variable Size, the previous approach was used in the same way, with just one minor adjustment, since Size is continuous, and therefore needs to be binned to obtain reasonable results when plotting. To do this we applied the Friedmann-Diaconis formula h = 2*IQR*n^1/3 to get number of bins = (max(x) – min(x)) / h. With this we proceeded to plot as usual and encountered a very different distribution of probabilities as before. (see fig. 39)

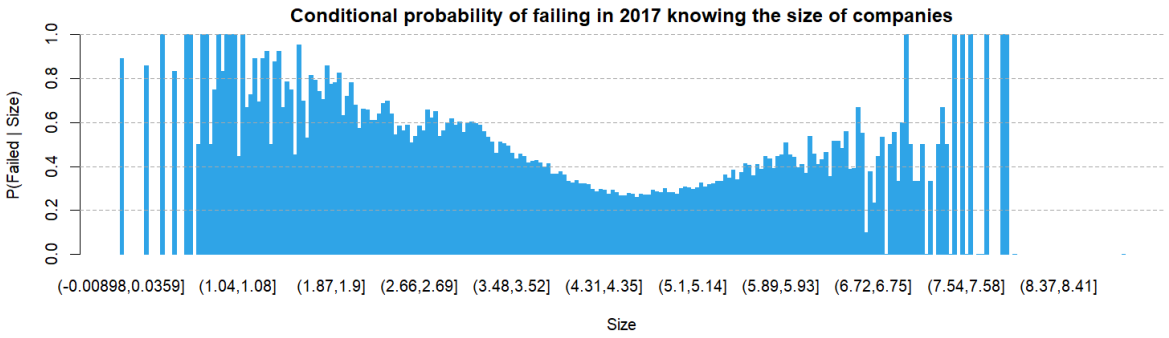
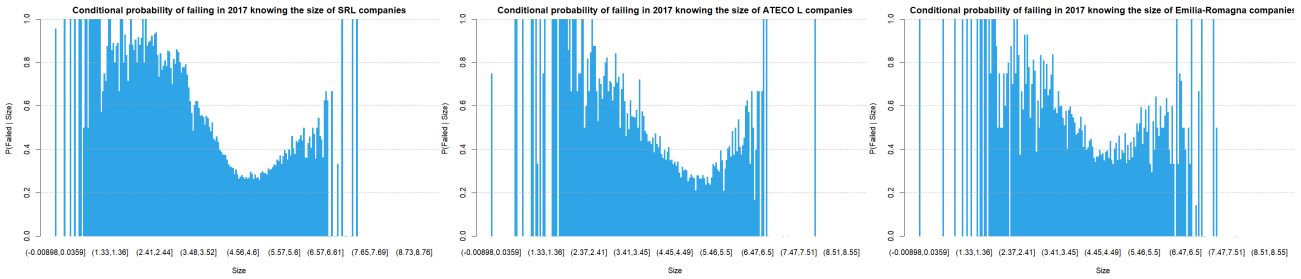


Figure 39 - Probability of failing in 2017 knowing the size of companies

Now we can observe a convex trend, where probabilities decrease when approaching the middle of the bins and then start returning upwards almost to the same level, however, at the last bins, the same problem as before appears again, rendering the last probabilities either 1 or 0. In any way, this confirms that the probability of failure conditional to certain variables is indeed different, based on this visual approach.

4.2.1 For specific company forms, ATECO Sectors and locations

As before, we look for changes in some different types of companies, these results can be seen in figure 40.



A) S.R.L. *B) ATECO L* *C) Emilia-Romagna*

Figure 40 - Conditional probability of failing in 2017 knowing the size of the company and another attribute

Overall, the same convex trend continues, although with different spans, remaining however the very high and low probabilities where the data is more sparse, this conforms to the previous analyses in which the distribution of Size has been shown to be normal, therefore having sparser tails in comparison to the dense middle.

4.3 Probability of failure conditional to Liquidity

Finally, we take a look at the probability of failure depending on the variable LiquidityRatio, again using the same binning technique as before. Here we can see the opposite trend as that shown in the Age section, where there was an initial increase in probabilities followed by a relative stability, in this case the probabilities start higher and then begin to decrease, then reaching a somewhat stable trend and finally

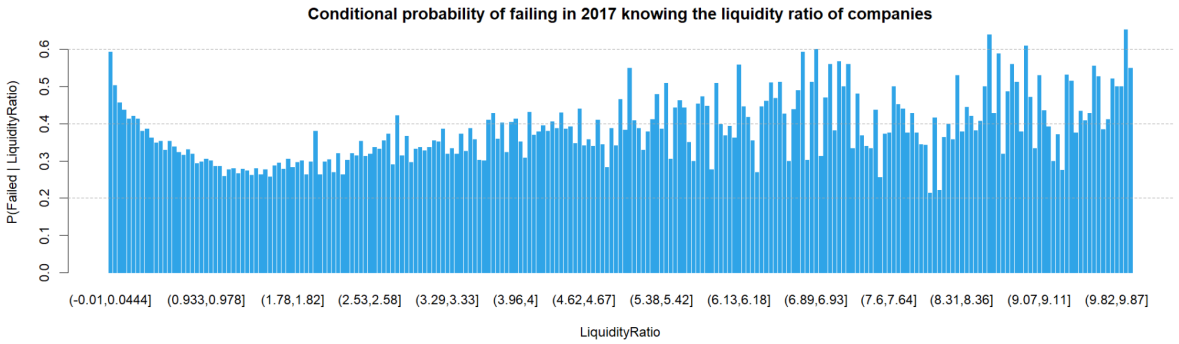
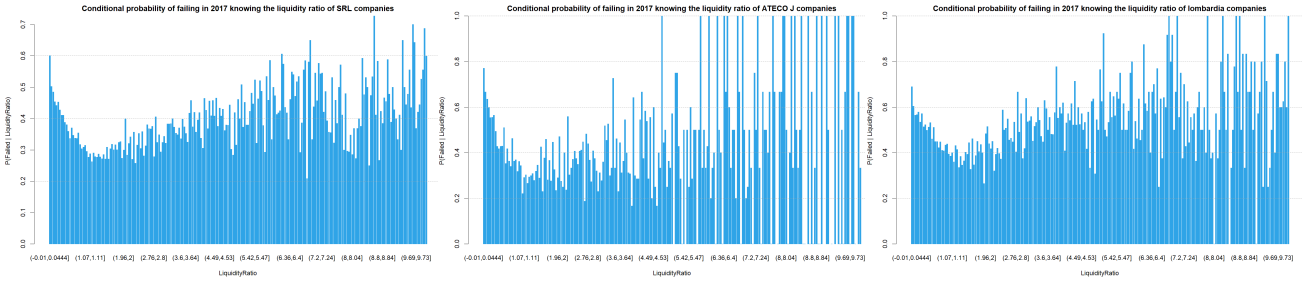


Figure 41 - Probability of failing in 2017 knowing the liquidity ratio of companies

4.3.1 For specific company forms, ATECO Sectors and locations

As in the previous exercises, this results can be further down broken by filtering by their geographical location, ATECO sector, or company legal form. Again, the general trend continues, not changing much between types of company. (see fig. 42.)



A) S.R.L. *B) ATECO J* *C) Lombardia*

Figure 42 - Conditional probability of failing in 2017 knowing the liquidity ratio of the company and another attribute

5 Question D

Regarding question D, the goal is to fit a parametric model to predict failure. For this we have decided to use logistic regression, which seems as a good fit to predict likelihood for our binary target variable.

The data used for this model was split temporally and non-numerical and irrelevant features were dropped as well as observations containing missing values. The chosen years were 2014, 2015, 2016 and 2017, chosen for their representability, providing a sufficient amount of observations and a somewhat balanced target class with 3,387 records pertaining to failed companies and 4,458 to active ones. After this, the data was split into a training and testing set making use of the *caret* library and considering a *holdout* method, in which we randomly selected 70% of the data for training and the rest for testing. After this, our training set consisted of 2,371 records pertaining to failed companies and 3,121 to active ones.

5.1 Developing a scoring and rating logistic regression model

Having the data ready, as a first approach we built a logistic model with the help of the R built-in function Generalized Linear Models glm, specifying the binomial family of models to go along with our needs, and then feeding the previous model into a *predict* function against the observations in the testing set, this is turn provides a probability scoring indicating the likelihood that each record belongs to a certain class. With these probabilities, a rating model is feasible, considering each observation from the testing set to belong to a failed company is the probability predicted was > .5, active otherwise. With this, it was possible to build a confusion matrix and obtain some performance measures, such as accuracy, precision, recall and F1-score. These performance scores can be fully appreciated in table 14, under the name of “Raw data - Base model”, along with the performances of further tried approaches, which aim to improve this baseline model.

Additionally, a ROC curve with its area under the curve and a calibration plot have been used to measure the performance of the model. From this first approach, we can see in fig. 43 that the model reaches a high false positive rate before achieving high true positive rates, indicating high rates of misclassification where the model predicted a failed company but instead it was an active company. As for the calibration plot in fig. 44, we can check if the model is well-calibrated, i.e. the probabilities effectively reflect the true likelihood of the event of interest and is therefore a reliable predictive model. Overall, the model seems to be well-calibrated, with an equal forecast along the diagonal, except for the first and last segments, in which some under and over forecasting can be seen.

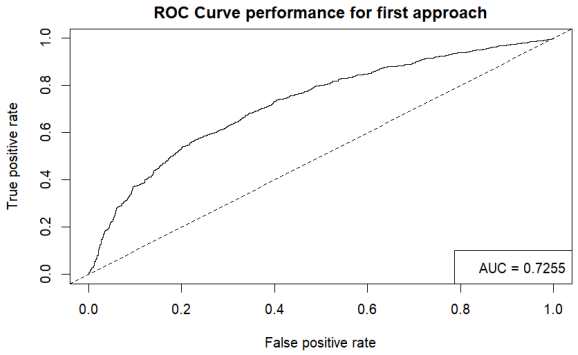


Figure 43 - Roc Curve for 1st approach

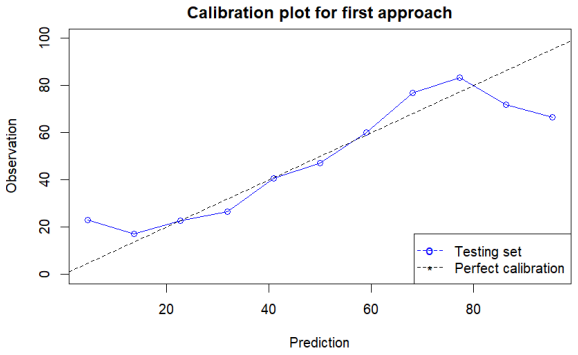


Figure 44 - Calibration plot for 1st approach

5.1.1 Solving multicollinearity

In order to try and improve the previous approach, we analyzed the correlation between independent variables, to see if multicollinearity might be hampering our model. To do this, we took a look at the correlation values of all features with the help of the function *findCorrelation* from library *caret* and set a threshold of correlation > .7 to find the most heavily correlated ones. This method found 44 variables with correlation greater than .7, however, an interesting thing to note here is that every variable found, also included the historical data "Year - 1" and "Year - 2", indicating that there might be very little difference between the different available years of variables. This led us to try and see the performance of the model by using only the last available year for each variable, expecting little change with respect to the original model since these features are all so correlated.

Overall, our intuition was correct, since the new model presents more or less the same performance values, except for the positive increase in recall, going from .7853 to .8569, the full results can be seen in table 14, under the name of "Selected Features 1".
Another positive improvement can be seen in the calibration plot of fig. 45, which got even closer to the diagonal, although with a sudden decrease under the diagonal in the last segments.

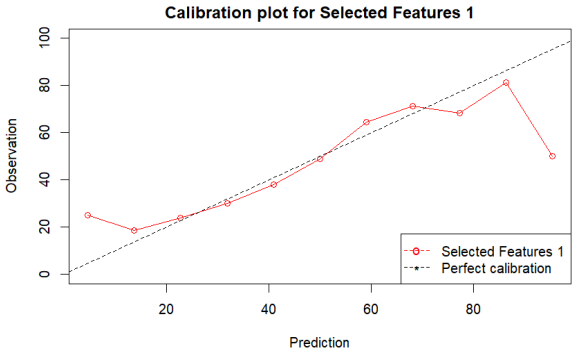


Figure 45 - Calibration plot for Selected features

These positive improvements inspired us to continue across this path, now introducing another correlation metric, VIF, variance inflation factor, a measure of the amount of multicollinearity in a set of multiple regression variables, equal to the ratio of overall variance to the variance of a model that includes only that single independent variable. Using the function *vif* from the library *car*, and removing any variable having vif > 5, we ended up removing the following variables, along those with cor > .7, as seen in table 13.

Thus, a third model was implemented, this time removing features contributing to multicollinearity, however, despite removing 9 more variables an ending up with only 18 features from the original 73, very little changed in terms of performance, AUC and calibration plot, proving that these final features have the greatest amount of significance and importance when building a logistic regression model to predict failure for the aida dataset. In any way, these results can be found under the name of "Selected Features 2" for comparison with the other models.

VIF & Correlation values of discarded features

Feature (VIF > 5)	VIF	Feature (Correlation > .7)	Correlation
Net working capitalth EURLast avail. y	365.92	Number of employeesLast avail. yr	.99
Net financial positionth EURLast avail. yr	344.66	Liquidity ratioLast avail. yr	.99
Total assetsth EURLast avail. yr	136.83	Total assetsth EURLast avail. yr	.96
Cash Flowth EURLast avail. yr	96.50	Cash Flowth EURLast avail. yr	.90
EBITDAth EURLast avail. yr	15.51	Return on sales (ROS)%Last avail. yr	.88
Number of employeesLast avail. yr	14.15	Net financial positionth EURLast avail. yr	.84
		Return on asset (ROA)%Last avail. yr	.82
		LeverageLast avail. yr	.73

Table 13 - VIF & Correlation values of discarded features

5.1.2 Solving class imbalance

Another important thing to consider when building a classifying model is the balance of the target variable within the training set, that is, the ratio of each of the class possibilities to total amount of available observations, For our particular training set, the dependent variable is overall balanced, with 43.17% of the observations belonging to failed companies and 56.83% for active ones. From this we can gather that our previous models were at least better than a random classifier, which could achieve .57 accuracy by predicting every record in the testing set as 1.

Now, even though this is not such a profound imbalance, we still decided to rebalance the data in order to further improve and analyze the results of a rebalancing technique, the chosen technique is an oversampler from the library *caret*, which balanced the class to an equal 50-50. After using this new data to feed the model, the model was again built and analyzed, sadly, no significant improvement can be seen, and some instances like the calibration plot actually worsened. These results can be seen at the end of the section under the name of "Rebalanced data".

A final approach was tested, combining the features obtained in "Feature Selection 2" and also applying class rebalancing. This model obtained the highest precision score of them all but also the lowest accuracy, indicating that there might be a systemic bias that lowers the overall trueness of the model. With also a low recall as well as poor calibration performance, we can state that this model is less reliable than the other tested before.

5.1.3 Model evaluation

Performance of different models

	Accuracy	Precision	Recall	F1 Score	ROC AUC
Raw data - Base model	.6834	.6963	.7853	.7381	.7255
Selected Features 1	.6753	.6904	.8569	.7647	.6858
Selected Features 2	.6754	.6900	.8586	.7651	.6874
Rebalanced Data	.6685	.7285	.6642	.6948	.7245
Combining methods	.6385	.7352	.6453	.6874	.6859

Table 14 - Performance of different models

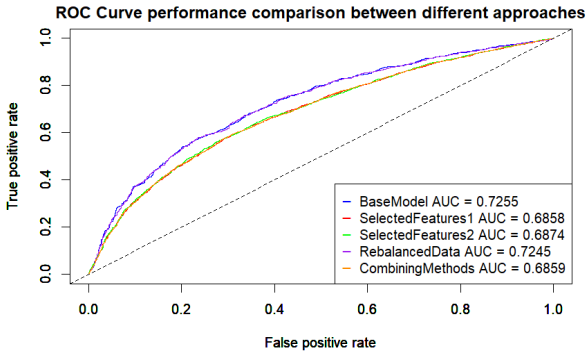


Figure 46 - Roc Curve comparison for all models

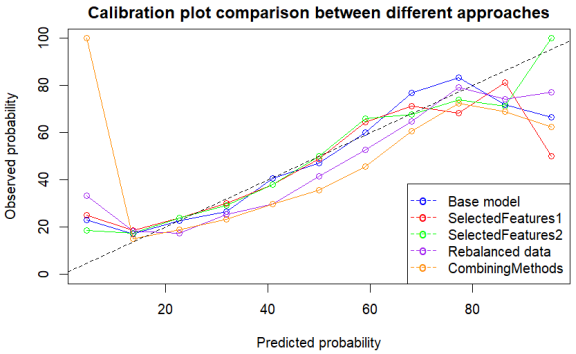


Figure 47 - Calibration plot comparison for all models

By looking at fig 47, we can see that the model with the highest AUC from the ROC curve is the base model of the 1st approach, followed by the model with rebalanced data, showing a consistently lower true positive rate TPR for the models with feature selection. As for the calibration plots in fig 46, we can see a trend of over-forecasting with the models using rebalancing, while those with feature selection tend to stay closer to the diagonal. Based on this and the overall performance seen in table 14, we could argue that the model with selected features from the correlation and VIF analysis provides the best results overall and is therefore considered the preferred method.

Additionally, we also observe the density function of both the fitted values and residuals for the different approaches tested before. On these values we used the F-test to compare the variance of the base model against the rest of the models, t-test for the means and Wilcoxon test to know whether the samples come from the same distribution.

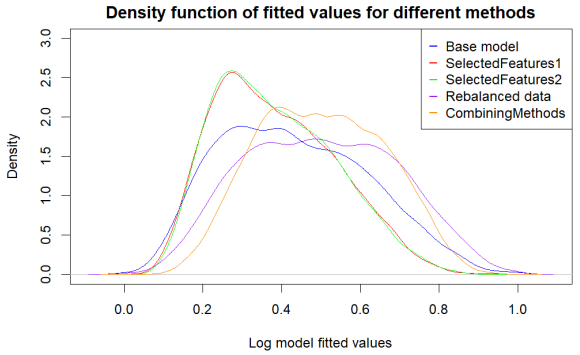


Figure 48 - Density of fitted values among models

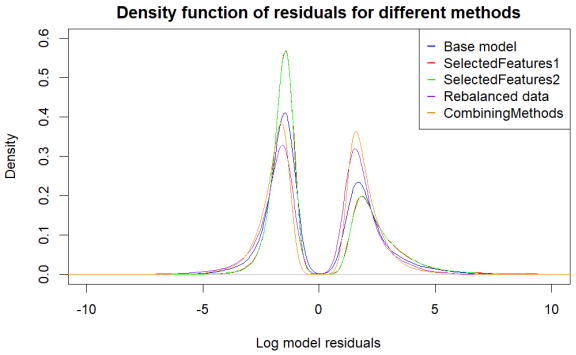


Figure 49 - Density of residuals among models

First the tests were performed using the fitted values shown in fig. 48, these results are presented in the table 15, where the only null hypothesis accepted is that against RebalancedData, confirming that there is indeed not any significant difference between the variances of the fitted values of the base model and that with rebalanced data.

	Test	F	p-value
SelectedFeatures1	f-test	1.4736	2.2e-16
SelectedFeatures2	f-test	1.5163	2.2e-16
RebalancedData	f-test	0.96072	0.126
CombiningMethods	f-test	1.3696	2.2e-16

Table 15 - Results obtained from f-test between variance of base model and the rest

On the other hand, the results for the t-test are presented in table 16. In this case, all the comparisons present the same p-value, below significance level and therefore rejecting the null hypothesis indicating they have the same mean.

	Test	p-value
SelectedFeatures1	t-test	2.2e-16
SelectedFeatures2	t-test	2.2e-16
RebalancedData	t-test	2.2e-16
CombiningMethods	t-test	2.2e-16

Table 16 - Results obtained from t-test between base model and the rest

In table 17, we can see the results of the evaluation by the use of the Wilcoxon test. In this case, the results are congruous amongst them and with the results presented in table 16.

	Test	p-value
SelectedFeatures1	Wilcoxon test	2.2e-16
SelectedFeatures2	Wilcoxon test	2.2e-16
RebalancedData	Wilcoxon test	2.2e-16
CombiningMethods	Wilcoxon test	2.2e-16

Table 17 - Results obtained from Wilcoxon test between base model and the rest

These tests were also applied to the residuals shown in fig 48, this time the f-test provided the same results, although with different p-values, with the t-test however, every p-value was above the significance level, therefore accepting the hypothesis saying that the residual distribution of all models share the same mean. Finally, for the wilcoxon test on residuals, the first two models got accepted null hypothesis, while the last two got rejected.

5.1.4 Rating model alternative

Since the previous models all had a fixed threshold $t > .5$ to be considered a failed company we now ask what would the results be if this threshold were to change. Following inspiration from the Bond Credit Rating used in investments, one could determine the quality of a company based on its probability to fail, in this way, companies that still got classified as active even when the threshold was set to $t > .1$ are companies that less than 10% probabilities to fail, therefore making them optimal candidates to investment for example. Moving this threshold, we end up with the following ratings and the amount of companies from the AIDA dataset belonging to it.

Rating predictions

	Rating	Predicted Failed companies	Predicted Active companies
$t > .1$	AAA	2334	19
$t > .2$	AA	2120	233
$t > .3$	A	1717	636
$t > .4$	B	1271	1082
$t > .5$	C	845	1508

Table 18 - Rating predictions

Meaning that if an investor wanted to invest in a certain company and used our model, and the model told them that the company falls under rating AA, the company would have at least 80% probability of remaining active, therefore guiding the decision of the investor.

6 Question E

6.1 Confidence intervals of predictions in a logistic regression model

As a last section of this work, we turn to the question of using confidence intervals for the predictions made by the model developed during section 5. To do this we used what was deemed as the best model during the previous section, that is, the logistic regression model with selected features after removing those with VIF > 5 and correlation > .7. With this same model and the same training and testing data, we now predicted the scores on the logit scale with the use of the basic predict function, this time with the argument *type = "link"*, which returns the log-odds ratios, instead of on the probabilistic scores as we did before with argument *type = "response"*, with this, we can now calculate a confidence interval for each log-odds ratio on the logit scale.

Since the log-odds ratio βj is estimated using Maximum Likelihood Estimation, the theory tells us that it is asymptotically normal and as such, we can use wald's large sample confidence interval, Wald's CI = βj ± z*σ(βj), where z is the z-score corresponding to the desired confidence interval and σ is the standard deviation of βj, we use z = 1.96 since we want a standard 95% confidence interval. After computing Wald's CI, all that is left to do is to take this interval back to the probabilistic scale, this is done with the help of the invariance property, which allows us to exponentiate to get e^βj ± z*σ(βj), which is a confidence interval on the odds ratio. Finally, with the endpoint transformation approach we apply a logit transform exp(βj) / (1 - exp(βj)) that takes us back to the probabilistic scale with our original prediction and its 95% confidence interval. Having both a lower and upper bound for our 95% confidence interval, we can plot a small sample of points belonging to the predictions of some of our testing data, showing the predicted probability along with the lower and upper bound of the confidence interval. This lets us visually appreciate the varying intervals among predictions (see fig. 50).

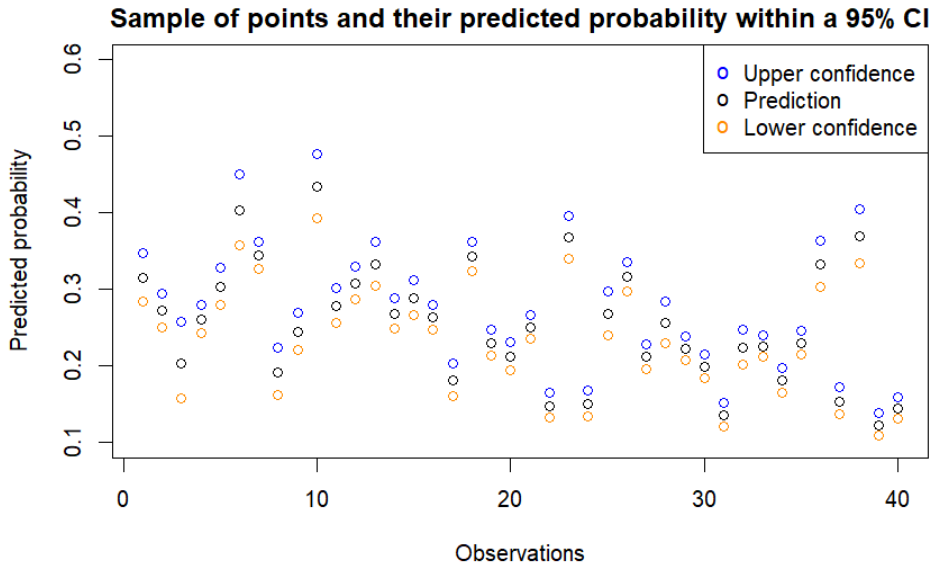


Figure 50 - Sample of points and their predicted probability within a 95% CI

In addition to this manual computation of the confidence intervals, we also made use of the library *modelbased*, which provides estimates and uncertainty for statistical models. After feeding it with our training and testing data, we get back a dataframe containing for each testing record, a prediction, a low confidence interval and a high confidence interval, these values along with the plot match those obtained with the previous method, confirming its validity.

Finally we implement yet another visual analysis, namely, confidence bands, with which we can observe how the confidence intervals shift along the values for a single variable within a logistic regression model, in this case we have chosen the variable *Age* and applied the same methods as before, as a result we get the fitted values of the model, that is, the predictions, and along with it, the lower and upper band of their confidence interval, as we can see in fig. 51, the confidence bands begin to widen when the age nears the 40 mark, this is consistent with the decrease in density we have seen in section 2 when plotting the density function of the *Age* distribution, indicating how the predictions have a closer confidence interval when having a higher density of data.

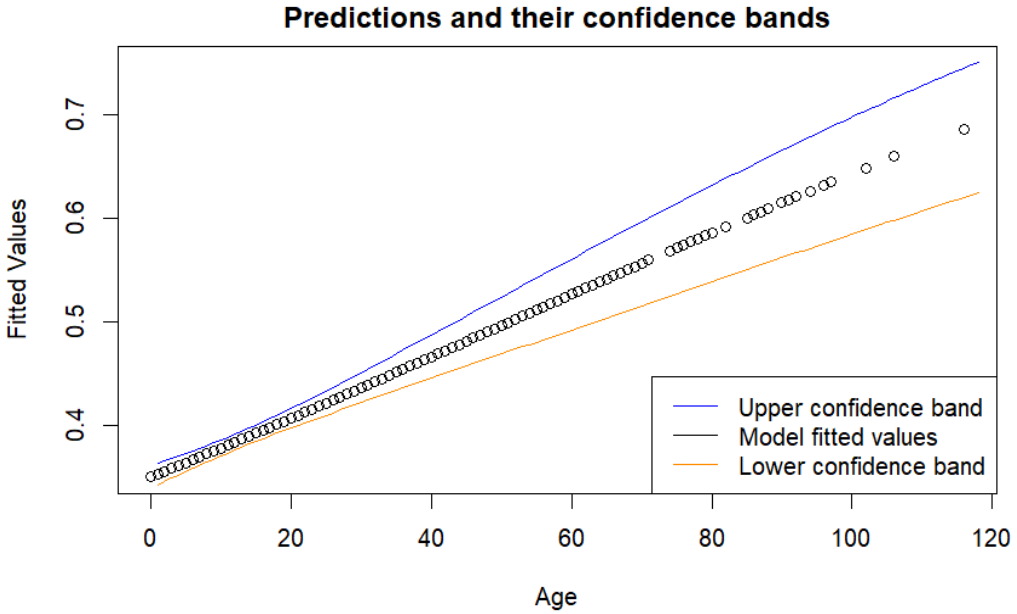


Figure 51 - Univariate model prediction with confidence bands

7 Conclusions

Summarizing the development of this project, in the Data Understanding & Preparation phase, the first stage consisted in getting to know the database, visualizing the effects of different attributes, development of functions and, most importantly, the definition of failure for this project. After creating the variables needed to perform the analyses required for this project.
We first evaluated how the distribution of age and size changes in a manner statistically significant between failed and active companies in a given year. The analysis carried out in 2018, showed that on average active companies tend to be more young than failed ones. Instead the size of the active companies tends to be higher than non-operating companies.
The objective of the second question, on the other hand, was to evaluate their distribution variables of the companies with a failure status in a certain period of time. Tests carried out on samples of the 2017 and 2018 reported that companies that went bankrupt in 2017 are on average younger than those that failed in 2018.
The third request requires analyzing, in a given year, the conditional probability of failures knowing the age, size of the companies. In between and almost, we encountered a decreasing trend in the probability of failure with increasing age and size. All of these have also been analyzed in different ATECO sectors, company legal forms, and the regions where companies were registered.
The following section pertains to the development of the failure prediction method and the evaluation of this model by using the methods of ROC and AUC seen in class.
Finally we explore how the implementation of confidence intervals allow to shift along the values for a single variable within a logistic regression model,

References

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. *The British Accounting Review*, *38*, 63–93. doi:10.1016/j.bar.2005.09.001

Dalgaard, P. (2008). Introductory Statistics with R. New York: Springer. ISBN: 978-0-387-79053-4

Dekking, Frederik & Kraaikamp, Cor & Lopuhaä, Hendrik & Meester, Ludolf. (2005). A modern introduction to probability and statistics. Understanding why and how. 10.1007/1-84628-168-7.

Moscatelli, M., Narizzano, S., Parlapiano, F., & Viggiano, G. (2019, Diciembre). Corporate default forecasting with machine learning. In *Temi di Discusione* (Issue 1256). Banca D'Italia.