Statistics for Data Science Project

Federico Canepuzzi,Gloria Segurini

Data Science and Business Informatics f.canepuzzi@studenti.unipi.it

Data Science and Business Informatics g.segurini@studenti.unipi.it

Statistics for Data Science, Academic Year: 2021/2022

Date: 05/09/2022

Abstract

The current report shows and explains the analysis of the AIDA dataset, which aims at a business failure prediction, risk factors exploration, and distributions investigation.

1 Introduction

The analysis regards the AIDA dataset. It presents itself with 80 features concerning 1894412 companies that compose the data frame. Most of these features are about economical and financial indicators of the companies during the last three years of activity. Some other features are present and regard information such as the geographical headquarter, opening year, last accounting closing year and legal form. Furthermore, the feature *Legal status* shows the financial situation of the firm at its last accounting closing date and it is the one used in order to assign a target value *Failed* equal to 1 or 0, indicating respectively failure or not. Indeed, the goals of this study are both the analysis of the characteristics of failing companies and business failure prediction. In order to perform such an analysis, the report presents a distinct section for every Task from A to E.

2 Overview and Methods

The first three Tasks require comparing the distributions and conditional probability of failure of some features such as *Age* or *Size* and detecting any change depending on the target value Failed, company form, *ATECO* industry sector, Last accounting closing year of the company and location (*Area*). The last two Tasks ask to develop a classification model to predict a company's probability of failure and analyze the results.

2.1 Code

The project was developed by using R programming language and R Studio IDE. Any required library such as DescTools, dplyr, ggplot2 and caret is indicated in the code scripts. The code scripts address Tasks A, B, C, D and E. The first three Tasks' scripts are saved as "Q%_data", where the % character changes based on the Task. For Task D there are three distinct scripts: QD_preprocessing, QD_feature selection and QD_E_classification, where the latter addresses both Tasks D and E.

2.2 Report structure

The current report has a distinct section for every Task from A to E. Concerning the Preprocessing procedures, a preliminary one was performed for Tasks A to C, while a more in-depth one was necessary for the Parametric and Machine Learning Models construction. Therefore, section 3 shows the preprocessing for the first three Tasks, which is the basis of Tasks D and E's Preprocessing shown in section 7.1.

3 Preliminary Preprocessing

Initially, it was necessary to find the definition of failure of a company. In the feature *Legal status* there are different values. The final choice was to set as *Active* only the companies already reporting the value "Active" in *Legal status*, while the ones reporting any declination of "Dissolved", "Bankruptcy" or "Liquidation" values were set to *Failed*.

The few records containing the values "Active (default of payments)" and "Active (receivership)" were deleted; this was done in order not to have any value coming from a failure-facing company among those considered active.

The records containing the values "Dissolved merger" or "Dissolved demerger" were also deleted, as a merge or demerge of a company does not provide sufficient information about the failure.

Therefore, the final dataset presents 65% of active companies and 35% of failed, respectively detected by the value 0 in the new column *Failed* and 1.

Afterwards, two new features were added: *Age* and *Size*.

- Age
  It was created by subtracting the values of *Last accounting closing year* and *Incorporation year*, that is between the year of publication of the financial statements and the year of the beginning of the company's activity. Since the total number of records was more than sufficient, all those presenting NA on at least one of the two features were deleted, as well as those presenting a negative value for *Age*.

- Size
  This feature was created by calculating the average number of employees in the three available years.

Considering the high number of companies with 0 or 1 employees, the final division was set in table 1

Size	Single	Micro	Small	Medium	Large	Extra Large
N. of Employees	0	[1,3]	[3,6]	[6,10]	[10,25]	[25 +]

Table 1: Relation between Size and Employees

Subsequently, many categorical attributes presenting many distinct values were modified, so as to reduce their quantity.

The attribute concerning the geographical region of the company's seat was modified by dividing the country (Italy) into four regions: "Sud", "Centro", "Nord-Ovest" and "Nord-Est".

The attribute *Legal Form* presented some values with very few records; these were joined under the value "Others". In particular, the following values were joined: "Association","Foreign company","Foundation""Mutual aid society", "Public agency", "S.A.P.A.", "S.N.C.", "S.A.S.".

Furthermore, the following legal forms were united under the value "S.C.A.R.L." :"S.C.A.R.I.", "S.C.A.R.L.P.A.", "Social cooperative company".

Finally, the feature *Ateco 2007 code* presented numerical values with six digits, with the first two indicating the macrocategory of the classification of the reference economical asset. Such values were therefore substituted with a letter indicating the macrocategory. In this case too, the categories presented too few records and were joined under the value "Others".

4 Task A

In order to perform this task, it was necessary to initialy choose a year for the analysis. By comparing the distribution of the active and failed companies throughout the various years, it is possible to notice how, before 2005, only records of failed companies are available. From 2005 on, the records presenting value 0 on the feature *Failed* start increasing, which overcome the number of failed companies in 2017 and become exponentially grater in 2018. In order to keep a balanced analysis, the year 2016 was selected, since it has a similar number of records for active and failed companies.

- *Age*
  Once the data for the year 2016 were selected, it was initially attempted to find the distribution of *Age* starting from the "Cullen and Frey graph". As shown in figure 1, the observed data and bootstrap samples indicate a possible Pareto or Gamma distribution. Through the Maximum Likelihood Estimator they were estimated the parameters for both distributions and it was performed a Kolmogorov-Smirnov (KF) test in order to test the distance from the assumed distribution. In both tests, the hypotheses were rejected, as they had a very low pvalue.

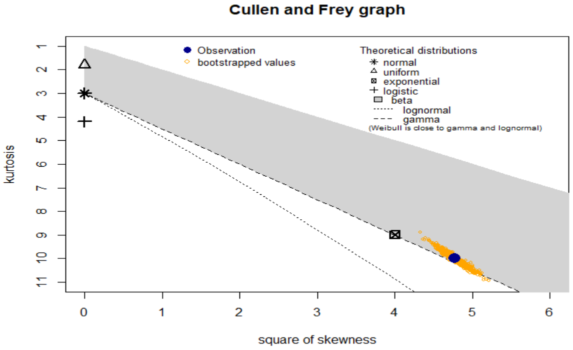

Figure 1: Cullen and Frey graph Age

Subsequently, the dataset was divided between the failed and active companies. In figure 2 it is shown the density of the estimation for *Age* for both *Failed* values. Despite the curves' differences, it was attempted to see if they had the same mean: it was impossible to use the Shapiro test for verifying the normality of the distribution, because of the high number of data (5000 records are permitted at most). Therefore, it was used the z-test (the t test was also usable alternatively). The hypothesis was rejected with a very low pvalue. The difference between the two averages has confidence intervals of 95% ranging from 2.7 to 3, pointing out that the failed companies are older than active ones on average.

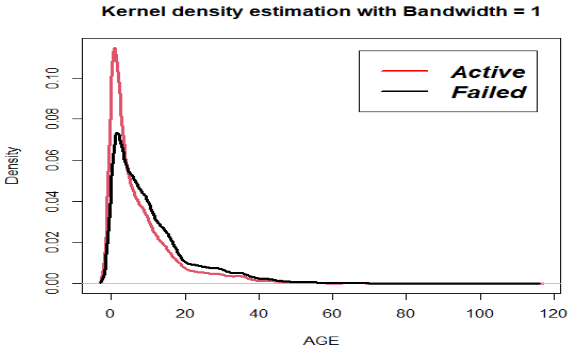

Figure 2: Density curve

As requested, it was analyzed whether there were any statistical differences or not, by fixing the attribute *Legal form*'s values. Since many records were available, the z-score was used once again, this time combined with the Bonferroni correction. The results are reported in table 2. It is visible that the hypothesis for active and failed companies with *Legal form* equal to "Others" to have the same mean on the *Age* attribute is not rejectable. The same can be said for companies having *Legal form* equal to "S.P.A.". The companies having other *Legal form* values have very low p-values and are all rejected. It is worth to evidence that the value "Consortium" presents a difference between the means with 99.2% confidence, ranging from -9.3 to -5.4 and this strongly deviates from the previously seen general value. This indicates that, in this case, active companies are older than failed ones. The same analysis was performed by fixing the values on the basis of the macrocategory of the ATECO industrial sector. In table 3 the test results are reported, conducted once again through a z-test and with the Bonferroni correction. This time, the hypothesis of same mean were all rejected, presenting very low p-values.

AGE - 2016				
	Failed	**Active**	**Difference**	**Statistical significance**
Legal Form	Mean	Mean	99.3% CI	p-value
S.R.L.	11.81	9.14	[2.44 ; 3]	<2.2e-16
S.C.A.R.L.	11.09	9.18	[1.15 ; 2.67]	1.576e-11
S.R.L. one-person	12.03	10.34	[1.11 ; 2.27]	3.043e-15
S.P.A.	25.19	23.59	[-3.03 ; 6.22]	0.353
S.R.L. simplified	1.26	1.01	[0.2 ; 0.3]	<2.2e-16
Other	13.9	15.1	[-3.84 ; 1.35]	0.1978
Consortium	12.65	20	[-9.3 ; -5.37]	<2.2e-16

Table 2: Z-test between distribution of Age in 2016 with a fixed Legal Form

AGE - 2016				
	Failed	**Active**	**Difference**	**Statistical significance**
ATECO	Mean	Mean	99.5% CI	p-value (alpha =)
G	10.01	6.56	[3.04 ; 3.87]]	<2.2e-16
C	14.07	9.05	[4.24; 5.80]	<2e-16
I	6.4	5.2	[0.67 ;1.73]	2.004e-10
F	12.46	8.74	[3.21 ; 4.20]	<2.2e-16
N	8.17	5.77	[1.79 ; 3.03]	<2.2e-16
Others	9.5	8	[0.95 ; 2.13]	2.22e-13
H	9.18	6.4	[1.87 ; 3.71]	<2.2e-16
J	9.29	7.78	[1.2 ; 2.82]	2.904e-12
L	16.5	15.4	[0.23 ; 1.91]	0.0003418
M	9.1	7.00	[1.51 ; 2.66]	<2.2e-16

Table 3: Z-test between distribution of Age in 2016 with a fixed Ateco value

- *Size*
  Since it deals with a categorical attribute with 6 distinct values, they were transformed into discrete values ranging from 0 to 5. A statistical test was conducted in order to check if the distribution of *Size* was the same for active and failed companies. It was used the Pearson's Chi-squared test, which rejected the hypothesis. It was therefore checked if the mean was the same for failed and active companies; once again, the z-test rejected the hypothesis. It is possible to claim that the means' difference between failed and active

companies ranges from -0.11 to -0.07 with a 95% confidence, suggesting a small shift of the active companies towards a bigger size. As with *Age*, it was checked whether the distributions changed on the basis of the *Legal form* values. For every distinct value it was performed a Bonferroni-corrected z-test where the H0 hypothesis states that the mean of the two distributions is the same; the results are shown in table 4. From the results it is not possible to reject the hypothesis for active and failed companies presenting *Legal form* equal to "S.C.A.R.L.", "S.P.A.", "Other" or "Consortium" to have the same mean. The same test was finally performed by fixing the ATECO industrial sector value with values reported in table 5. For the categories G, C, and N it is not possible to reject the null hypothesis. For category H it is visible an increase of the mean for failed companies with difference confidence intervals ranging from 0.02 to 0.38 (99.5%).

SIZE - 2016				
	Failed	**Active**	**Difference**	**Statistical significance**
Legal Form	Mean	Mean	99.3% CI	p-value
S.R.L.	0.82	0.97	[-0.17 ; -0.12]]	<2.2e-16
S.C.A.R.L.	1.49	1.48	[-0.09 ; 0.11]	0.78
S.R.L. one-person	0.79	0.91	[-0.2 ; -0.05]	8.286e-06
S.P.A.	1.72	2.09	[-0.93 ; 0.18]	0.07
S.R.L. simplified	0.66	0.79	[-0.18 ; -0.08]	3.909e-11
Other	0.9	0.89	[-0.24 ; 0.27]	0.90
Consortium	0.33	0.3	[-0.09 ; 0.14]	0.60

Table 4: Z-test between distribution of Size in 2016 with a fixed Legal Form

SIZE - 2016				
	Failed	**Active**	**Difference**	**Statistical significance**
ATECO	Mean	Mean	99.5% CI	p-value
G	0.82	0.85	[-0.075 ; 0.017]]	0.07875
C	1.4	1.44	[-0.13; 0.05]	0.204
I	1.33	1.45	[-0.2 ;-0.03]	0.000111
F	0.72	0.87	[-0.2 ; -0.095]	<9.978e-15
N	1.21	1.25	[-0.17 ; 0.08]	0.3063
Others	0.77	1	[-0.3 ; -0.15]	2.22e-16
H	2.02	1.81	[0.02 ; 0.4]	0.0017
J	0.66	0.75	[-0.18 ; 0.003]	0.006451
L	0.18	0.2	[-0.07 ; 0.004]	0.01378
M	0.5	0.55	[-0.14 ; -0.004]	0.002936

Table 5: Z-test between distribution of Size in 2016 with a fixed Ateco value

5 Task B

In this Task, they were compared the distributions of *Age* and *Size* of the failed companies in two distinct years. The choice of the rears was based on the 2008 crisis: it was decided to compare years 2009 (the first following the crisis, and therefore strongly affected by the crisis) and 2016 (which was a long way from the crisis).

- *Age*
  In figure 3 it is reported the density curve for *Age* in the two years.

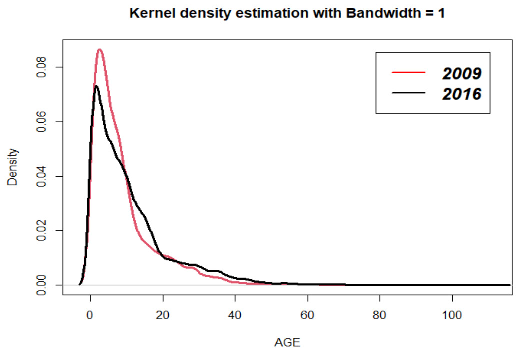

Figure 3: Density estimation

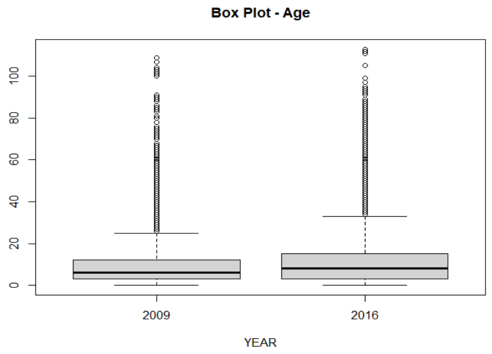

Figure 4: Boxplot on Age : 2009 vs 2016

The pvalue was obtained through the Kolmogorov-Smirnov test, which refuses the hypothesis that the two curves come from the same distribution. Also the z-score pvalue which checks for the same mean has very low values and refuses the hypothesis. The difference on average between 2009 and 2016 presents confidence intervals of 95% equivalent to -2.02 and -1.75, suggesting a higher age for the failed companies in 2016. Through the Cullen and Frey graph it was tried to figure out whether the data of 2009 and 2016 derived from a well-known distribution (figure 5 and 6).

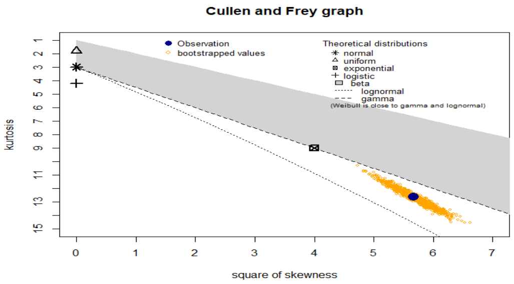

Figure 5: Cullen and Frey graph - 2009

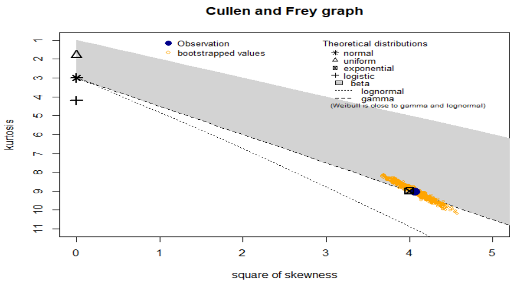

Figure 6: Cullen and Frey graph - 2016

From the graph, it seems the data of failed companies in 2009 are far away from such distributions. The data of the 2016 failed companies suggest the following possible distributions: exponential, Pareto, Gamma. The Kolmogorov-Smirnov tests performed after calculating the parameters with the Maximum Likelihood Estimation refuse all the three previously hypothesized distributions. In this case too, it was checked if the distributions change on the basis of the *Legal form* attribute. However, in this case it was necessary to delete from the analysis the values "S.R.L.", "Simplified" and "Others". The first was deleted, because such formed-companies only exist from 2012 and the second and third, because it presents very few records. Table shows the results of the z-test for the mean with Bonferroni correction. The only non-rejectable hypothesis for the mean is the "S.C.A.R.L." one (table 6).

The same analysis was also conductedby selecting the location value of the *Area* attribute. For all the four values the z-test rejects the same-mean hypothesis (table 7).

AGE - 2009,2016				
	Failed	**Active**	**Difference**	**Statistical significance**
Legal Form	Mean	Mean	99% CI	p-value
S.R.L.	11.82	9	[2.68 ; 3.11]	<2.2e-16
S.C.A.R.L.	11.09	11.23	[-0.83 ; 0.56]	0.6091
S.R.L. one-person	12.03	7.5	[4.11 ; 5]	2.2e-16
S.P.A.	25.19	18.2	[3.77 ; 10.19]	2.16e-8
Consortium	12.65	8.11	[3.29 ; 5.8]	<2.2e-16

Table 6: Z-test between distribution of Age in 2009 and 2016 with a fixed Legal Form

	AGE - 2009,2016			
	Failed	**Active**	**Difference**	**Statistical significance**
Area	Mean	Mean	98.75% CI	p-value
Sud	10.9	8.42	[2.15 ; 2.83]	<2.2e-16
Nord_Ovest	12.77	9.6	[2.82 ; 3.6]	<2.2e-16
Nord_Est	12.23	9.14	[2.7 ; 3.5]	<2.2e-16
Centro	11.6	9	[2.26 ; 3]	<2.2e-16

Table 7: Z-test between distribution of AGE in 2009 and 2016 with a fixed Area

- *Size*

As requested by the task, the *Size* distributions were also examined for failed companies in 2009 and 2016. In figure 7 it is reported the bar plot.

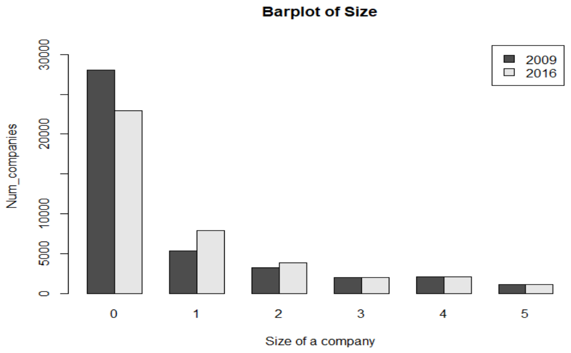

Figure 7: Bar plot of Size

In order to check if they had the same distribution on *Size*, it was performed the Pearson's Chi Squared test. The hypothesis is rejected with a very low pvalue. The z-test for testing the similarity of the means gets rejected. In table 8 they are reported the z-score results with the Bonferroni correction, fixing the *Legal status* value. All the equal-mean hypothesis are rejected. The same approach was followed for *Area*; in this case, we found out that for the value "Nord-Est" it is not possible to refuse the hypothesis of having the same mean for the two distributions (table 9).

SIZE - 2009, 2016				
	Failed	**Active**	**Difference**	**Statistical significance**
Legal Form	Mean	Mean	99% CI	p-value
S.R.L.	0.82	0.7	[0.1 ; 0.15]]	<2.2e-16
S.C.A.R.L.	1.49	0.85	[0.54 ; 0.73]	<2.2e-16
S.R.L. one-person	0.79	0.94	[-0.2 ; -0.09]	1.25e-09
S.P.A.	1.71	2.01	[-0.76 ; -0.015]	0.077248
Consortium	0.33	0.14	[0.084 ; 0.28]	1.72e-06

Table 8: Z-test between distribution of Size in 2009 and 2016 with a fixed Legal Form

	SIZE - 2009,2016			
	Failed	**Active**	**Difference**	**Statistical significance**
Area	Mean	Mean	98.75% CI	p-value
Sud	0.97	0.8	[0.11 ; 0.21]	<2.2e-16
Nord_Ovest	0.86	0.77	[0.05 ; 0.14]	8.017e-08
Nord_Est	0.8	0.77	[-0.01 ; 0.08]	0.05511
Centro	0.9	0.7	[0.16 ; 0.24]	<2.2e-16

Table 9: Z-test between distribution of SIZE in 2009 and 2016 with a fixed Area

6 Task C

In this Task it was requested to analyze the distribution of the probability conditioned to the failure of a company on *Age* and *Size* for a given year. Once again, it was chosen the year 2016, since it has a good balance between the number of failed and active companies.

- *Age*

Since *Age* is a continuous attribute, it was divided into bins, so as to compute the conditional probability for every bin. The division was done by using quantiles, to have a similar number of data in each bin. The bar plot with conditional probability of failure for the five bins is reported in figure 8.

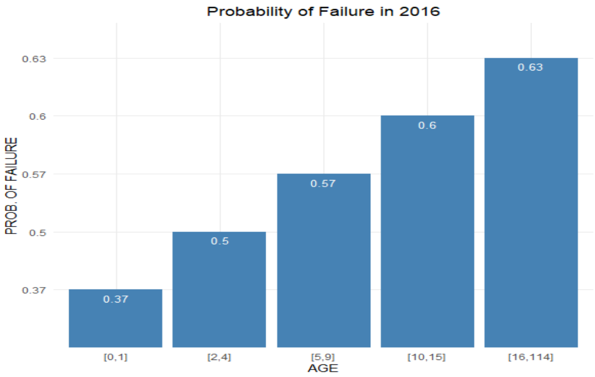

Figure 8: Probability of failure on Age Bins in 2016

From the graph it is visible how the conditional probability of failure grows as the company Age increases, reaching 63% for bin [16-114].

In order to check in the probability of failure on every age bin was equal to the probability of failure for the whole year 2016, a series of Bonferroni-corrected binomial tests was also performed. All the tests were rejected.

The binomial test was subsequently tested, this time failure for also fixing the values on *Legal form*, *Area* and *ATECO*. In all three cases, the various conditional probabilities were tested against the conditioned probability on *Age*, without fixing the latter features. Hypothesis H0 always claims that the conditional probabilities are equal. The results are shown in the three following tables 10, 11 and 12.

AGE BINS					
	[0,1]	[2,4]	[5,9]	[10,15]	[16,114]
LOCATION	p.value	p.value	p.value	p.value	p.value
SUD	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
CENTRO	0.001208	1.392e-09	2.534e-11	<2.2e-16	6.125e-10
NORD-EST	4.263e-15	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
NORD-OVET	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16

Table 10: Binomial test on Probability of failure of Age Bins with a specific Area vs Probability of failure of Age Bins in general

AGE BINS					
	[0,1]	[2,4]	[5,9]	[10,15]	[16,114]
LEGAL FORM	p.value	p.value	p.value	p.value	p.value
S.P.A.	0.001253	0.09874	0.0992	0.07762	0.003121
Other	5.438e-09	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16
Consortium	0.2525	0.02351	0.01707	0.07091	<2.2e-16
S.R.L. one-person	0.1722	<2.2e-16	4.65e-06	5.314e-06	7.597e-11
S.R.L. simplified	<2.2e-16	0.0001072	NULL	NULL	NULL
S.C.A.R.L.	0.0001283	3.8e-09	0.8836	0.3995	0.4708
S.R.L.	<2.2e-16	0.002284	0.8416	0.1066	2.257e-10

Table 11: Binomial test on Probability of failure of Age Bins with a specific Legal Form vs Probability of failure of Age Bins in general

AGE BINS					
	[0,1]	[2,4]	[5,9]	[10,15]	[16,114]
ATECO	p.value	p.value	p.value	p.value	p.value
C	0.0007406	0.189	8.933e-07	1.941e-06	<2.2e-16
F	<2.2e-16	1.725e-11	4.895e-06	0.001849	0.244
G	0.2948	0.1222	6.089e-05	0.001128	2.165e-09
H	0.126	0.5034	0.3156	0.1682	0.9611
I	0.000202	6.751e-11	0.0007197	1.34e-10	4.375e-12
J	8.816e-16	8.683e-07	0.004909	0.0172	0.01165
L	1	0.001559	0.001717	0.02664	<2.2e-16
M	1.798e-11	2.003e-12	4.169e-09	5.372e-05	7.227e-06
N	4.598e-05	0.1165	0.1446	0.4079	0.0008334
OTHERS	0.4926	0.2662	0.004474	0.02425	1.145e-10

Table 12: Binomial test on Probability of failure of Age Bins with a specific Ateco vs Probability of failure of Age Bins in general

With respect to *Area*, all tests concerning a fixed value for *Area* reject H0. however, by fixing *Legal Form* and *Ateco* some hypotheses cannot be rejected. In particular, in *Legal Form* the values”S.C.A.R.L.” and ”S.P.A.” present non-rejectable p-values by conditioning on three out of five age bins. In *ATECO* the value ”H” actually presents high p-values for every bin: it is not possible to ever refuse the hypothesis that the conditioned probability is equal to the one obtained by not conditioning on a specific *ATECO* value.

- *Size*

As for *Age* it was analyzed the conditioned probability based on *Size* for companies in 2016. Barplot in figure 9 shows the various probabilities.

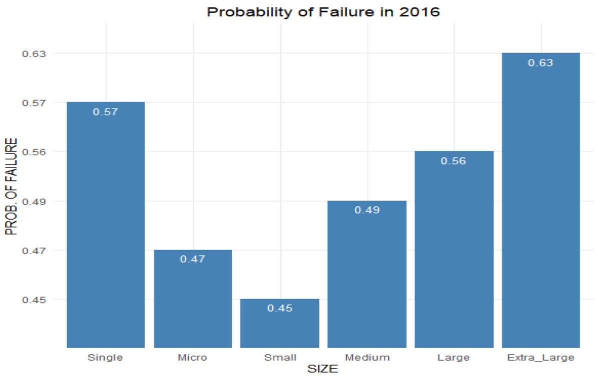

Figure 9: Probability of failure on Size in 2016

It is observable how the failure probability is higher for ”Single” and ”ExtraLarge” companies. Also in this case, the binomial test was performed between a company failure probability based on *Size* and the general 2016 failure probability. All the hypotheses to have the same value were rejected.

It was also tested to fix the values on *Legal form*, *Area* and *ATECO* and perform again the binomial tests, where H0 is equal to having the same failure probability without fixing the previous values. The results are reported in table 13, 14 and 15.

SIZE						
	SINGLE	MICRO	SMALL	MEDIUM	LARGE	EXTRA LARGE
AREA	p.value	p.value	p.value	p.value	p.value	p.value
SUD	<2.2e-16	<2.2e-16	<2.2e-16	2.743e-10	3.094e-06	9.771e-06
CENTRO	<2.2e-16	2.342e-06	3.686e-07	5.815e-05	0.01512	0.7534
NORD-EST	<2.2e-16	<2.2e-16	5.358e-12	1.574e-05	0.003145	0.05016
NORD-OVET	<2.2e-16	<2.2e-16	<2.2e-16	2.598e-15	9.185e-06	0.00524

Table 13: Binomial test on Probability of failure of Size with a specific Area vs Probability of failure of Size in general

SIZE						
	SINGLE	MICRO	SMALL	MEDIUM	LARGE	EXTRA LARGE
LEGAL FORM	p.value	p.value	p.value	p.value	p.value	p.value
S.P.A.	1.318e-08	0.001076	0.04028	0.1452	1	0.2001
Other	<2.2e-16	<2.2e-16	<2.2e-16	6.584e-11	1.287e-10	6.169e-06
Consortium	3.198e-10	0.1176	0.3747	0.6489	0.6818	0.407
S.R.L. one-person	<2.2e-16	2.319e-14	3.588e-08	0.0008439	0.0006375	0.02452
S.R.L. simplified	<2.2e-16	<2.2e-16	<2.2e-16	<2.2e-16	1.128e-09	1.323e-05
S.C.A.R.L.	0.004555	0.01157	0.2273	0.7973	0.3272	0.2219
S.R.L.	<2.2e-16	<2.2e-16	9.97e-07	0.009156	0.4459	0.8858

Table 14: Binomial test on Probability of failure of Size with a specific Legal form vs Probability of failure of Size in general

SIZE						
	SINGLE	MICRO	SMALL	MEDIUM	LARGE	EXTRA LARGE
ATECO	p.value	p.value	p.value	p.value	p.value	p.value
C	<2.2e-16	1.203e-08	4.418e-06	0.0001484	3.838e-09	0.004969
F	0.2374	<2.2e-16	5.95e-08	0.2088	0.8168	0.4852
G	0.6734	2.05e-08	1.538e-07	0.000583	0.08204	0.6684
H	0.4685	0.0005254	0.4271	0.01033	0.4539	0.7941
I	<2.2e-16	1.791e-05	2.272e-09	1.947e-12	5.909e-08	0.05195
J	1.883e-10	2.821e-05	0.006281	0.0007734	0.8397	0.01041
L	5.773e-06	0.06327	0.4049	0.7959	0.2479	0.8188
M	3.419e-16	4.105e-08	6.845e-05	0.3947	0.2127	0.1923
N	0.09218	0.0674	1	0.7233	0.5625	0.5653
OTHERS	8.515e-05	0.5478	0.0007689	6.708e-06	1.194e-09	0.002218

Table 15: Binomial test on Probability of failure of Size with a specific Ateco vs Probability of failure of Size in general

From these three table it is possible to observe as for the values Consortium e S.C.A.R.L. in Legal Form there are some high pvalues. The same thing is observed in Ateco for the category H, L and N. In Area all the H0 hypotesis are rejected.

7 Task D

7.1 In-depth Preprocessing

The first step was to remove irrelevant features, such as *Tax Code Number*, *Company name* and *File*. Subsequently, since in every financial indicator only the values concerning the last three years were reported, all the companies younger than three were deleted. Leaving them would have involved several missing values. It was then decided to exploit the values of every economical indicator to compute two derived columns. For every index, the average in the three years (two years or even just one in case of missing values) and the trend were calculated. The latter takes the difference between the value of the last year and the value of the two previous years (one in case of missing values) and is normalized by dividing by the mean. By analyzing the data at hand, it was decided to only keep years from 2007 to 2018, since the records were unbalanced in the other ones. Furthermore, all the features presenting more than 50% missing values were deleted. Otherwise, the analysis could have been incorrect.

More tests were also performed in order to check the independence of the variables with respect to the target. Since the target is a discrete value, the continuous features were discretized into bins to perform the Chi Square test. It was not possible to reject the hypothesis for the attribute *Total assets turnover (times)_trend*, which was then eliminated.

Then, multicollinearity was checked. Such a phenomenon can cause problems with logistic regression. After rescaling the data, it was then realized a model with logistic regression, which permitted the analysis of Variance Inflation Rate. *Cash Flowth EUR_mean* presented a much higher value with respect to the other features and was therefore eliminated.

Finally, to finish feature selection it was runned the Akaike Information Criterion algorithm, which eliminated 16 features.

At this point, the dataset was divided into training (TR) and test set (TS); the TR takes all the records ranging from 2007 to 2017 in the variables *Last accounting closing year*, while the TS takes the records with year 2018. The data was scaled on the basis of the values of the TR and outliers were eliminated only on TR, in order not to influence the TS. The features *Age* and *Current liabilities/Tot ass.%_trend* were eliminated, as they presented too many outliers.

Finally, all the rows still presenting missing values were dropped and were balanced on the basis of the target value in TR and TS, by randomly eliminating some records of the target-dominant value.

The final result is a TS composed by 88126 records equally divided between active and failed companies with 22 features and a test set with the same number of features and 28880 balanced records.

7.2 Logistic Regression

At this point it was fit a Linear Regression Model by using the library caret. It was performed a 10-Fold Cross Validation repeated 5 times. The chosen scoring metric was the AUC.

COEFFICIENTS	Estimate	Std. Error
(Intercept)	76.04753	7.87144
Cash Flowth EUR_mean	-433.86489	49.95201
Current liabilities/Tot ass.%_mean	2.14634	0.08731
Current ratio_mean	0.88717	0.11323
EBITDA/Vendite%_mean	-3.32183	0.22185
Interest/Turnover (%)%_mean	0.43093	0.11042
Leverage_mean	2.49336	0.59017
Liquidity ratio_mean	0.75959	0.12018
Net financial positionth EUR_mean	66.83668	42.38245
Number of employees_mean	436.69347	50.92963
Profit (loss)th EUR_mean	-142.95797	36.21179
Return on asset (ROA)%_mean	-5.98936	0.36110
Return on equity (ROE)%_mean	-0.09139	0.06986
Return on sales (ROS)%_mean	-0.80479	0.05863
Solvency ratio (%)%_mean	-0.86904	0.05927
Total assets turnover (times)_mean	1.00450	0.04619
Interest/Turnover (%)%_trend	0.34815	0.03012
Liquidity ratio_trend	-0.08779	0.05420
Total assetsth EUR_trend	-4.04397	0.06512
AreaCentro	0.04362	0.01759
AreaNord_Est	0.58536	0.02044
AreaNord_Ovest	0.67797	0.01948
ATECO_CATF	-0.31417	0.02512
ATECO_CATG	-0.21976	0.02408
ATECO_CATH	-0.30546	0.03836
ATECO_CATI	-0.67778	0.03295
ATECO_CATJ	-0.22310	0.03688
ATECO_CATL	-0.44379	0.03263
ATECO_CATM	-0.17066	0.03203
ATECO_CATN	-0.31797	0.03436
ATECO_CATOthers	-0.50303	0.02970
Legal form‘Other‘	-2.22915	0.08429
Legal form‘S.C.A.R.L.‘	-0.01949	0.06287
Legal form‘S.P.A.‘	0.51846	0.10146
Legal form‘S.R.L.‘	-0.02129	0.05896
Legal form‘S.R.L. one-person‘	0.13272	0.06103
Legal form‘S.R.L. simplified‘	-0.91977	0.07449

Table 16: Logistic regression Coefficients estimation

The estimated coefficients values are reported in table 16. Since the coefficient of the logistic regression represent the log odds ratios, a positive one shows that as the value of the independent variable increases, the mean of the dependent variable also tends to increase and viceversa. It is clear how some coefficients present high values. In particular, the feature *Number of employees_means* presents a coefficient with value 436.69, which is visibly large; this indicates that the feature weighs much more than the others.

7.2.1 Test set

In table 17 is reported the confusion matrix obtained by passing the TS to the model. The value of accuracy is equal to 0.65 (with confidence interval at 95% equal to 0.6418 and 0.6528).

It is also reported in figure 10 the calibration curve that shows the no perfect calibration of the model. It tends to assign a probability of failure greater than expected, in the range [0.2, 0.6] e lower in the range [0.9, 1].

	Reference	
Prediction	Active	Failed
Active	8699	4445
Failed	5741	9995

Table 17: Confusion Matrix Logistic Regression on test set

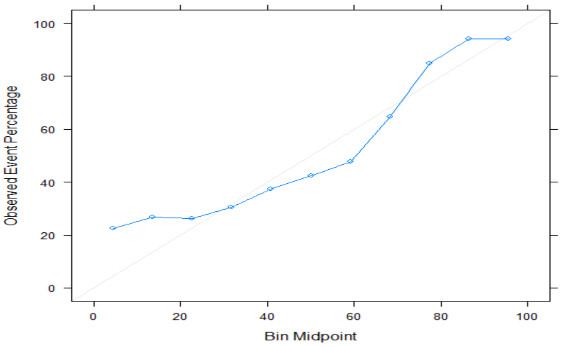

Figure 10: Calibration Logistic Regression

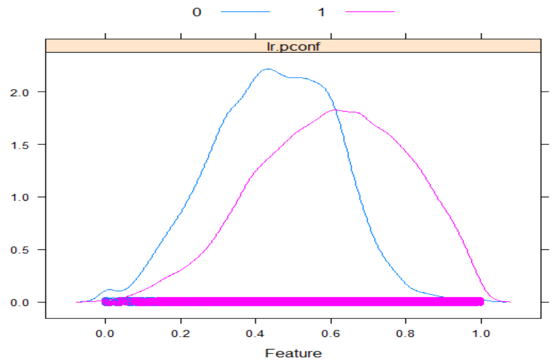

Figure 11: Probability density of failure between Failed (Class) 1 and Active company (class 0)

It was then performed the Wilcoxon test, which lets us understand that on average, the predicted probabilities of failure are between 0.154 and 0.163 with a 95% confidence interval. From figure 11 the shift between the two classes is visible.

7.3 Random Forest

A Random Forest Model was constructed by using the same dataset as before. Also in this case, a 10 Fold Cross-Validation was performed 5 times. The grid-search approach was also performed in order to set the parameter for the number of attributes to select, testing values from 5 to 25 every 5. For the model selection 20 trees were used.

In table 18 it is reported the confusion matrix computed on the TS with Random Forest. The accuracy value is 0.69 with 95%confidence intervals between 0.686 and 0.698.

	Reference	
Prediction	Active	Failed
Active	10597	5071
Failed	3843	9369

Table 18: Confusion Matrix Of Random Forest

In figure 12 it is reported the calibration curve performed on the TS. The improvement with respect the the same curve shown for the logistic regression is evident. Also the Wilcoxon test underlines the improvement: now the distance on the mean between the probability of failure of the 2 class is in the range [0.20006, 0.24999]with 95% of confidence (figure 13). In figure 14 is showed the Roc Curve of the Random forest.

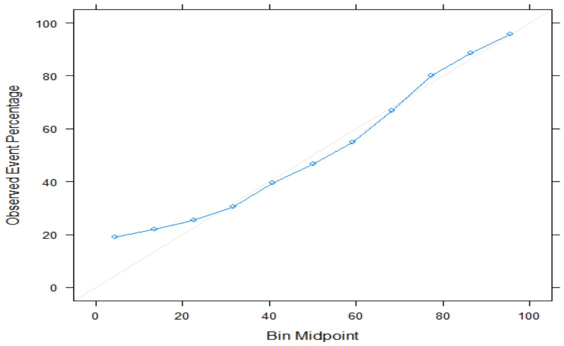

Figure 12: calibration Random Forest

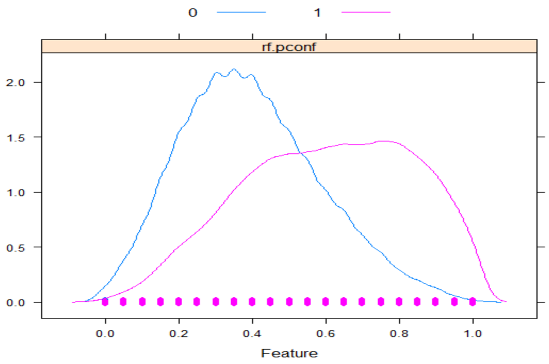

Figure 13: Probability density of failure between Failed (Class) 1 and Active company (class 0)

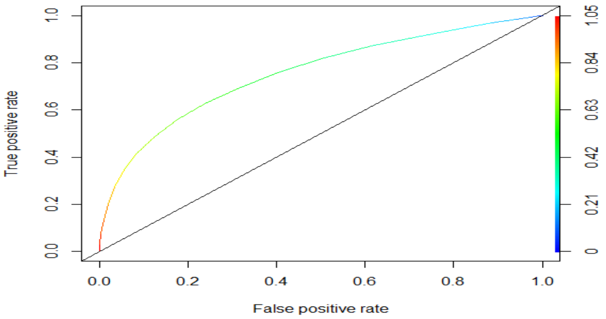

Figure 14: Roc Curve Rrandom Forest

7.4 Model Comparison

In order to compare the two classification models, it is created a dataframe with the values of AUC got from the previous analysis. For each classificator there are 50 records. The box plot in figure 15 show the quantile of the data.

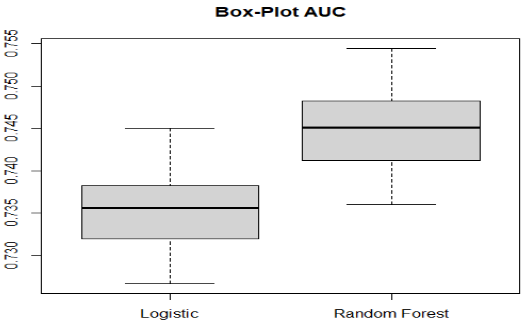

Figure 15: Box Plot Logistic Regression and Random Forest

In order to execute a statistical test to compare the models, it was tested the normality with the Shapiro test. In both the cases it is not possibile reject the H0 hypothesis. So a t-test was applied with hypothesis of same mean: the pvalue very low suggest to reject H0.

7.5 Rating

Besides the companies probability of failure prediction thanks to the logistic regressor, it was also decided to set a label on the basis of the default risk. Starting from range [0,0.1] to 1 increasing by 0.1, 10 labels from A to J were assigned. The final model shows that the companies with rating equal to A are the most secure ones, while those with value J face a default risk.

It was also addressed a binomial test with the Bonferroni correction for every rating value. The TS was divided on the basis of the rating and the H0 hypothesis was tested: it claims that the true probability in every group is less than or equal to the upper limit of the rating class.

Rating	**Threshold**	**P-VALUE**
A	0.1	1
B	0.2	1
C	0.3	1
D	0.4	1
E	0.5	1
F	0.6	1
G	0.7	1
H	0.8	1
I	0.9	1
J	1	1

8 Task E

Task E required to study a selective classification for the logistic regression model realized in Task D. For this purpose it was created a function which tests the best constraining-values for the model to abstain the value prediction.

In figure 16 it is possible to see on the y axis the error calculated from the accuracy and on the x axis the coverage of our TS. Indeed, by abstaining from some uncertain predictions, the coverage of the predicted data.

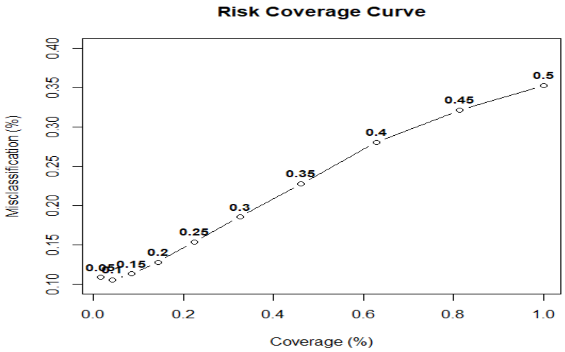

Figure 16: Error - coverage curve