RISK OF BUSINESS FAILURE

Giovanni Scognamiglio, Martina Trigilia
g.scognamiglio2@studenti.unipi.it, m.trigilia@studenti.unipi.it
Data Science & Business Informatics
Statistics For Data Science 628PP, Academic Year: 2021/22

Abstract

In the following project we used applied statistical methods and machine learning models to analyse and predict business failure. With the data at hand, we statistically assessed the superior discriminatory power of non-parametric models compared to parametric ones in predicting default risk. We also evaluated the marginal gain in classification accuracy that non-parametric models obtain compared to parametric models when applying a selective binary classification.

1 Method

The project was performed using R and the R Studio workspace. External libraries used include DescTools for descriptive statistics, Caret for training and evaluating the predictive models, Ggstatplots for statistical tests and Ggplot for plotting.
For the predictive modelling tasks, the Area under the Receiver Operating Characteristic (AUC-ROC) measure was used as scoring quality measure and Accuracy was used for measuring classification performance. The data was temporally split in a training and test set. The training set was re-balanced to correct for covariate drift. Model's hyperparameters, when applicable, were selected through grid-search optimization using cross-validation. The models were then retrained on the entire training set with the best parameters. The ROC curves and calibration plots were obtained by assessing the models' performances on the test set; whilst the repeated evaluation measurements for comparing scoring performance among models were drawn from each fold of a repeated cross-validation on the training set with the same random seed.

2 Data Understanding and Pre-processing

The objective of the project is to study the variables that influence a firm's probability of default. The dataset used is "Aida", a dataset containing historical financial indicators of italian firms, developed by Bureau Van Dijk, an analytics company. The original dataset contains slightly less than two million rows representing unique firms. It has over 80 variables referring to various financial indicators and firm's characteristics, many of which are also reported for the last two years prior to the last year the company submitted its balance sheet.

Defining the target variable
The first step was to define the target binary variable to identify whether a firm was considered "Failed" based on its legal status. Our approach was to find a solution that would both consider the class balance of the newly defined variable and the semantic meaning of the variable. We first removed the firms that had a "Dissolved (demerger)" 10.96 and "Dissolved (merger)" as status, given their low appearance (4.6e-3% and 2.2e-2% of tot. rows respectively) and the uncertainty in classifying such firms as purely failed or active. We labelled firms as "Active" if their legal status was one of the following: "Active", "Active (receivership)", "Active (default of payments)". The remaining firms were labelled as "Failed". The newly created binary target variable was name "Failed" and had 37% of failed companies and 64% active ones.
We then proceeded with feature engineering and defined the variables object of the various questions.

Defining variable Age
The variable "Age" was defined as the difference between the firm's last accounting closing date and its incorporation year. The age variable is to be considered as the firm's age in the year it last presented it's balance sheet.
We removed the firms that did not have an incorporation year (3.7e-3% of tot. rows). Firms that had a negative age (i.e. when its incorporation year post-dated the last accounting closing year) were also dropped (1.4e-5% of tot. rows).
The variable was also checked for any relationship to a known parametric distribution. We identified two potential distributions (Exponential and Pareto) and estimated their parameters with through Maximum Likelihood Estimation (MLE) (Figure 12 in Appendix). Although both distribution visually appear to well fit the data, a goodness of fit performed with the Kolmogorov-Smirnov (KS) test assessing the quality of the fit for both distributions rejected the hypothesis that the data originated from any these distributions (P-value = 2.2e-16 for both).

Defining variable Size
The variable Size was defined as the natural logarithm of the variable Total Asset (in

thousands).
We first removed statistical outliers. We then observed an unusually high number of firms who's total Asset equaled exactly 10, visible in Figure 1 A at Size=log(10)=2.3. The probability of default (PD) of these firms was much lower than the PD of firms with total assets in the neighbouring quantile range.[1] We thus deducted that the value 10 was a standard defaults imputation for a firm's total assets. We imputed these inconsistent values with values of total assets from an interval appropriate to their relative probability of failure.[2] In Figure 1 we can see the variable Size before and after the cleaning.
As with variable Age, an assessment of plausible parametric distributions fitting was performed. As we initially though the variable total asset to be a log-normal, we presumed that the variable Size, being the natural logarithm of the former, would be normally distributed. Although the normal distribution with expectation 12.25 and standard deviation 2.04 (derived with MLE) appears to be visually a good fit, we had to reject the hypothesis that Size originated from a normal distribution based on the KS-test results (P-value = 2.2e-16).

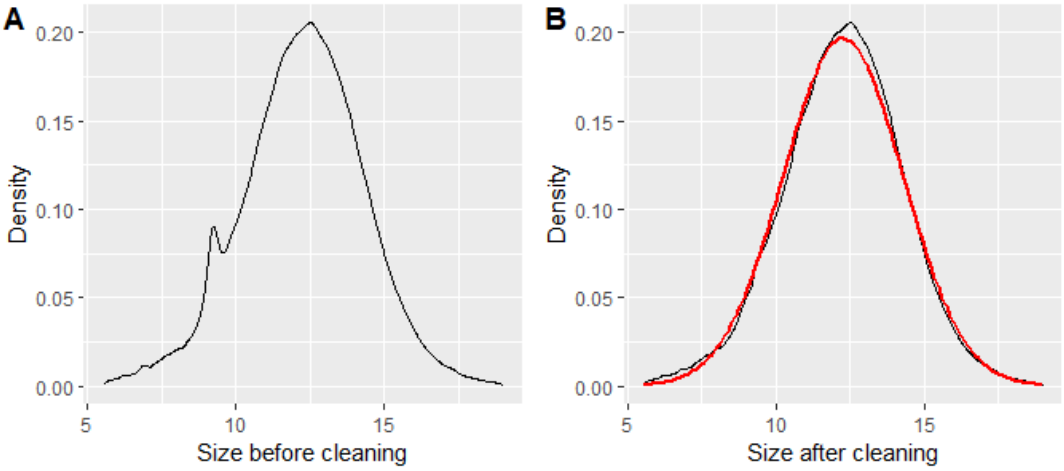

Figure 1: Distribution Densities of Size before and after cleaning process.

General pre-processing
In order to facilitate the understanding of the results of the various questions, we proceeded to strategically reduce the number of categories of the various variables object of

[1]The firms who's total assets were equal to 10 had a PD of 39%, which was much lower than the firms who's total assets values were between 9 and 12.6 (6th and 9th quantile respectively) which had an average PD of 53%.

[2]We identified a new range between 58 and 105 (25th and 35th quantile respectively) where firms had on average the same PD of the firms with tot asset equal to 10 and imputed the inconsistent values with randomly selected values from that range.

our study.
Concerning the variable "Location", we used an external source (Istat Italian District) for mapping the region of the firm with its geographical position (Nord-Est, Nord-Ovest, Centro, Sud, Isole).
Regarding the variable "Legal Form" we removed missing values and merged the following legal forms together: *S.A.P.A., Foundation, Foreign company, S.C.A.R.I., Public agency, Mutual aid society , Association , S.N.C. , S.A.S. , S.P.A.* in a global category named "Other" as they contained few rows. Moreover, the legal forms *Social Cooperative company, S.C.A.R.L.P.A.* were merged into *S.C.A.R.L.* because they all represent a type of cooperative company. In Figure 2 we can observe how the target variable is distributed within various Legal Forms.
The variable "ATECO.2007code" was transformed by mapping it to its relative industry sector, according to Istat Classification, resulting in a new factor variable named "Ateco Name" with 21 levels. We then checked the conditional probability of Failed over Ateco Name and merged together the industry sectors whose conditional probabilities were similar; the conditional probabilities on the original Ateco levels is reported in Figure 13 in the Appendix. The merged values of the variable Ateco Name are shown in Table 9 in Appendix. A bar plot showing the frequency of each Ateco Name is reported in Figure 14 in Appendix.

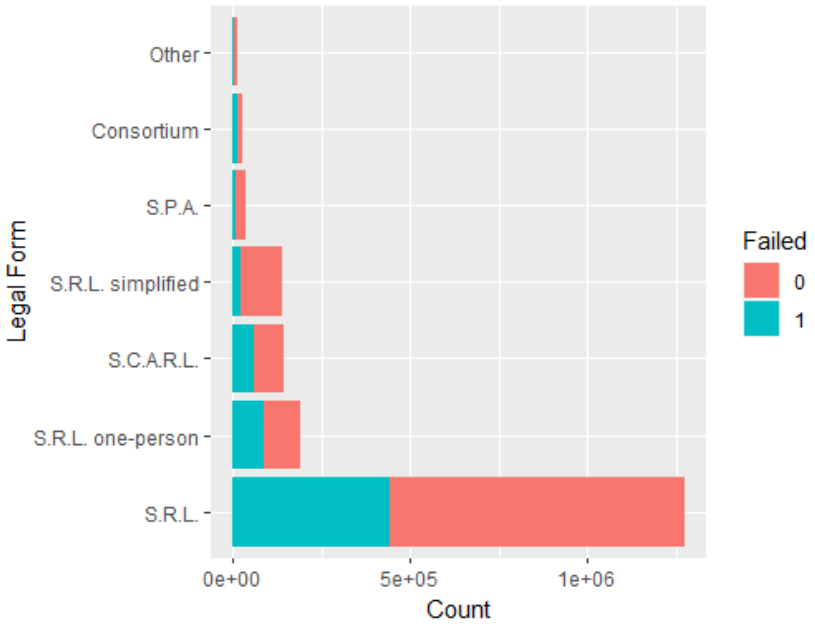

Figure 2: Bar Plot of Legal Form after pre-processing with comparison of Failed and Active firms

3 Questions

3.1 Question A

Compare the distributions of size and age between failed and active companies at a specific year.
In order to select a specific year, we considered the distribution of the dependent variable "Failed" together with the number of firms with respect to every year.
We saw that that from 1990 until 2004, there are very few firms and all of them are failed. From 2005 until 2013, the number of firms increases while the probability of failing is stable around 75%. Starting from 2015, the probability of failing drops steadily while the number of firms gradually increases. For 2017, there are 131'000 firms in the dataset and the probability of failing is 34%. In 2018, the number of firms peaks to 900'000 firms and probability of failure shrinks to 8%.
We selected the year 2016 as it provided the most balanced data subset with respect to failure rate (probability failure was 53%) and still provided a large number of firms for our analysis (around 80'000 firms).

Size
We first considered variable Size. We first made a KS test between the distribution of Size of firms failed against active in 2016 (P-value = 2.2 e-16). We rejected the null hypothesis that the two samples are drawn from the same distribution. We then proceeded to compare the means of the samples. We first checked the normality of the sample assumption with the Shapiro test and rejected the hypothesis that both were drawn from a normal distribution (P-value: 0.015 for failed companies; P-value: 4.63e-05 for active companies).
The samples of Size for failed and active companies are both large (greater than 22,000 rows) and therefore, with the assumption of general data, we performed a t-test, which returns a p-value of 2.2e-16, for this we can reject the null hypothesis that the true difference in means for the samples is equal to 0, and we can state that with 95% confidence the difference in mean of size between active and failed is between 0.66 and 0.73, with mean estimate of 11.80 and 11.35 respectively.
In Figure 3 (A) we can see the density plot of Size for failed and active companies.

Next, as required by the question, we checked whether there is a statistically significant difference between the distributions of Size between failed and active firms for a specific legal form and for a specific industry sector. With the assumption of large general data, we performed multiple t-tests with p-values corrected with Bonferroni, taking into account that all treated samples had large size ($\gg 50$).

Table 1 summarizes the size differences between active and failed with respect to Legal

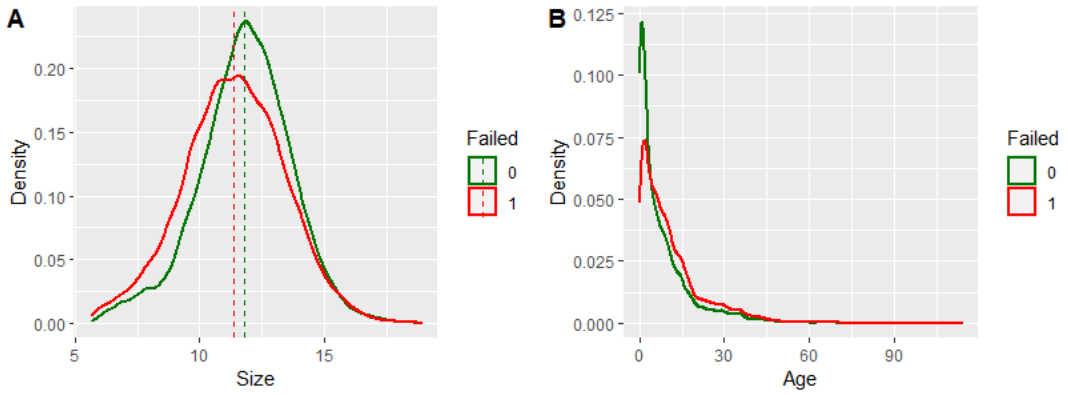

Figure 3: Size and Age distribution for failed and active firms

Form: the second and third columns show the means, while the fourth column shows the 99.1% confidence intervals for the differences (initially set to 95% and subsequently corrected for multiple comparison with the Bonferroni correction). We can observe that for most company forms there is a statistically significant difference for the mean of size in active and failed companies. In particular, the mean of size is larger for active than for failed companies, except for companies whose industry sector is Other where we can not reject the hypothesis that active and failed companies have the same average size.

SIZE				
	Active	Failed	Difference	Statistical significance
Legal Form	Mean	Mean	99.3% CI	p-value (alpha = 0.007)
Consortium	12.07	10.90	(0.89 , 1.44)	7.1e-20
Other	12.78	13.62	(-0.04 , 0.56)	ns
S.C.A.R.L.	11.16	10.99	(0.04 , 0.30)	4.3e-4
S.P.A.	15.51	14.39	(0.54 , 1.68)	1.1e-7
S.R.L.	12.22	11.52	(0.66 , 0.74)	0
S.R.L. one-person	12.55	11.68	(0.76 , 0.97)	4.2e-20
S.R.L. simplified	10.29	9.73	(0.47 , 0.66)	1.8e-20

Table 1: T-test for a specific industry form. H0: true mean Size difference between active and failed companies.

Table 2 shows the comparison between size of active and failed firms with respect to ATECO names. The corrected confidence level used was 99.6%. The mean size of active firms was statistically higher for all Ateco categories except for H (Trasporti) and S (Altri

Servizi) where we can not reject that active firms have the same size as failed ones.
Concluding on the variable Size, we saw that, globally, we could reject the hypothesis that active firms had equal or smaller size than failed ones. Moreover, this difference remains most often the same even when conditioning on Legal Form or Ateco names where only in a few subcategories the difference resulted being not significantly different from zero.

SIZE				
	Active	Failed	Difference	Statistical significance
Ateco Name	Mean	Mean	99.6% CI	p-value (alpha = 0.004)
A	12.07	10.90	(0.55, 1.25)	7.2e-14
EL	12.78	13.62	(0.68 , 0.91)	1.7e-20
FN	11.16	10.99	(0.28 , 0.44)	7.9e-20
GM	12.22	11.52	(0.37 , 0.52)	4.1e-20
H	12.55	11.68	(-0.11 , 0.29)	ns
I	10.29	9.73	(0.37 , 0.60)	5.2e-20
JC	11.90	11.51	(0.28 , 0.49)	6.2e-20
KD	12.05	11.16	(0.56 , 1.22)	1.1e-14
PB	11.24	10.54	(0.32 , 1.10)	1.47e-7
Q	10.96	10.43	(0.21 , 0.83)	2.1e-6
R	11.34	10.69	(0.41 , 0.88)	2.6e-15
S	10.97	10.72	(-0.01, 0.52)	ns

Table 2: T-test for a specific Ateco Sector. H0: true mean Size difference between active and failed companies is equal to zero.

Age

We then considered the Age variable. Also in this case, the KS test refused the null hypothesis that the active and failed firms are drawn from the same distribution (P-value = 2.2e-16). In Figure 3 (B), we can see the plot comparison between the distribution of Age of failed and act Although the variable had a quasi exponential shape, we relied on the assumption of large data and the Central Limit Theorem to compared the means of active and failed firms with a t-test.
The p-value obtained was 2.2e-16 and thus we refused the null hypothesis that the true difference in means of age for both samples is equal to 0, and that with 95% confidence the difference in mean of size between active and failed is between -2.92 and -2.624, showing that mean of Age for failed companies is higher than for active companies, as opposed to the result obtained for Size.

From the analysis of how the difference of failed against active firms changes for a specific legal form (reported in Table 10 in Appendix) we see that the difference is inverted for Consortium companies, which has a positive difference, as opposed to the global difference which is negative. Other and S.P.A. companies do not have a an age difference between active and failed statistically significant. All other companies have a negative difference with a 99.3% confidence.

Concerning how the difference changes for a specific Ateco name (Table 11 Appendix) we see from the P-values and confidence interval that most companies have a statistically significant negative difference. Only for firms operating in "PB - Istruzione & Miniere" and "Q - Sanità" is the difference is not statistically different from zero.

In conclusion, we generally rejected that failed firms have an equal or higher average age then their active counterpart. The negative difference remains unchanged for most subcategories when conditioning on Legal Form and Ateco name. An exception can be seen in Consortium firms where, with a 95% confidence, active firm actually have a higher average age.

3.2 Question B

Compare the distributions of size and age of failed companies over different years

For the following task, we chose to analyse the difference of size and age of failed companies between years 2014 and 2016. We selected those on the basis of a balanced proportion of failed firms and a significant number of rows. We did not select two consecutive years so as to render the analysis more interesting. Only failed companies were considered as requested by the task.

Size
We first performed a KS-test between the distribution of size of failed companies in 2014 and in 2016 and reject the null hypothesis that the distribution are identical based (P-value = 6e-10). We graphically checked the the distribution of Size for both years and saw that they have the same shape, as shown in Figure 5 (A).

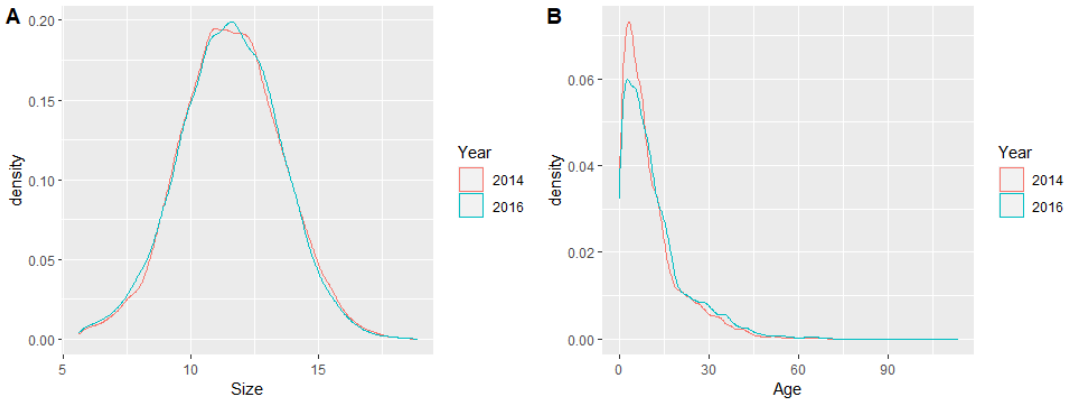

Figure 4: Distribution Densities of Size and Age for failed companies in 2014 and 2016.

To increase our confidence in asserting the distribution have the same shape, we removed the difference of the expectations of the two distribution from one distribution and proceeded again with the KS-test and obtained a p-value of 0.067. This tells us that the distribution are indeed related by a shift in location. We thus proceeded with the Wilcoxon rank-sum test. As expected, the Wilcoxon test provided a p-value close to zero and thus rejected the null hypothesis that the shift in location is zero. The estimated difference in location is 0.12 (95% CI 0.094 , 0.150).

Considering the distributions of Size in 2014 and 2016 conditioned to a specific legal form and a specific location, we proceeded controlling graphically the assumption of same shape. We found that for all legal forms and locations the distributions of both years have the same shape, and for this reason the comparison was made with the Wilcoxon

test.
In Table 3 and Table 4, we reported the results obtained for Legal Form and Location respectively. In this case, the significance level was adjusted to take into account also the comparison with the result of the Wilcoxon test for the general distributions (indicated with ALL* in the tables) between year 2014 and 2016.

SIZE			
	Shift (2014 - 2016)		Statistical significance
Legal Form	estimate	99.3% CI	p-value (alpha = 0.007)
ALL*	0.12	(0.08 , 0.16)	2.2e-16
Consortium	-0.30	(-0.60 , -0.02)	0.003
Other	-0.14	(-0.52, 0.24)	ns
S.C.A.R.L.	-0.30	(-0.16 , 0.10)	ns
S.P.A.	-0.02	(-0.50 , 0.46)	ns
S.R.L.	0.04	(-0.01 , 0.08)	ns
S.R.L. one-person	0.06	(-0.04 , 0.16)	ns
S.R.L. simplified	-0.18	(-0.33 , -0.02)	0.002

Table 3: Wilcoxon rank-sum test. H0: Size distributions of 2014 and 2016 considering a specific Legal Form are related by a shift in location, that is the difference of the two expectations, and this shift is zero

SIZE			
	Shift (2014 - 2016)		Statistical significance
Location	estimate	99.2% CI	p-value (alpha = 0.008)
ALL*	0.12	(0.08 , 0.16)	2.2e-16
Centro	0.15	(0.08 , 0.23)	2.2e-8
Isole	0.11	(-0.02, 0.25)	ns
Nord - Est	0.11	(0.02 , 0.20)	5e-4
Nord - Ovest	0.09	(-0.50 , 0.46)	ns
Sud	0.17	(0.09 , 0.26)	1.3e-8

Table 4: Wilcoxon rank-sum test. H0: Size distributions of 2014 and 2016 considering a specific location are related by a shift in location, that is the difference of the two expectations, and this shift is zero

When conditioning on Legal Form we see that Consortium and S.R.L. simplified firms both have a significant negative shift meaning the size of that firms' industry sector increased from 2014 to 2016, as opposed to the shift of the general distributions, which is

instead positive. For all other companies we could not reject that the shift was equal to zero thus meaning their size did not significantly vary from 2014 to 2016.
Regarding firm's size conditioned on Location the Table 4 tells us that for firms whose geographic location is 'Centro', 'Nord - Est' and 'Sud' we can reject the null hypothesis that the shift of location is zero. More precisely 99.2% confidence interval (CI) of the difference in shift for these firms is positive, as the one for the general distribution.

Age

As regards to the variable Age, we repeated the same procedure. First, we performed the KS-test that rejected the null hypothesis that the two distributions are equal. Next, we checked the assumption of same shape graphically, which indeed hold for the general distribution of age over the two years, see Figure 5. We further visually inspected whether the assumption of same shape holded aslo when conditioning on a firm's legal form or location. We noticed that for some industry sector types (S.R.L. simplified and Other) the assumption of same shape was not satisfied. For this task alone, we decided to exclude these two types of industry sector, so as to provide a more rigorous analysis of the shift in location.
The Wilcoxon test over the general distribution of 2014 and 2016 provided a very small p-value (p value: 4.44e-14) so we can assert that the shift in location is statistically significant from zero, even if it is very close to it (shift equaled 1.87e-5 with 95% CI between 4.10e-5 and 1.48e-5).
When conditioning on a firm's legal form (Table 12 in Appendix), we further corrected significance level with Bonferroni to take into account also the presence of the general distributions in the multiple tests (indicated as "ALL*" in the tables).
For most legal forms, we rejected that the shift in location between the age distributions of 2014 and 2016 is zero. With a 99.1% confidence we can state that the true shift for most legal forms is negative, in contrast to that of the general distributions shift, with the exception of S.C.A.R.L companies, whose shift in location is not statistically different from zero.
We found that for all locations the distributions of both years have the same shape and proceeded with the Wilcoxon test (Table 13 in Appendix). We rejected the null hypothesis for firms located in 'Nord-Est' and 'Nord-Ovest', this means that in these locations the firms which presented their last balance-sheet in 2016 were statistically older with respect to those who presented it in 2014. For all other firms' location, we can not reject the null hypothesis that the shift is equal to zero.

3.3 Question C

What is the probability of failure conditional to size and age of firms at a specific year?

We filtered our data set by the year 2016, as it contains a sufficient number of rows for our analysis and a good balance over the target variable.
The task required to derive the probability of failure conditional on a given size (age). This is given by the number of companies failed with a specific size(age) value divided by the total number of firms which have that specific value of size(age)
The task also required to calculate the probability of failure conditional on a given value of size(age) and a given value of Legal Form, Ateco Name and Location.
We realized that for the variable Age not all subgroups have enough failed firms to allow us to calculate the conditional probability. Specifically, we saw a higher concentration of companies with values for the variable Age ranging from 0 to 25, but the number of companies went down dramatically for higher values of Age. When further conditioning by company form/industry sector/location the problem is even more evident.
For this reason, we thought the best approach was to binnerize the variable Age in equal frequency bins, so as to have a larger number of firms for each bin and to make the comparison of conditional probabilities more meaningful and understandable.
We also binnerized the variable Size, only for this task, in order to calculate the conditional probability. The variable Size took the values of "Small," "Medium" and "Large" for ranges in [5.63,10.8), [10.81,12.5) and [12.46,18.9], respectively.

Size
Our first step was to find out whether the probability of failure conditioned on Size changes for different size values. Our approach was to apply the binomial test, which compares a sample proportion to a hypothesized proportion. Specifically we performed multiple binomial tests to do an all pairs comparison among the proportion of Failed when conditioned to the various Size categories. The significance level was set at 0.016 to account for the multiple tests and the alternative hypothesis was to greater for each test.
For all three tests the resulting p-value is 2.2e-16, so we reject the null hypotheses. In Figure 5 (A) is shown the bar plot of the probability of failure conditional on Size.

As required by the task, we now want to understand whether the probability of failure conditional on a specific size bin changes when we consider a specific type of firms, so we proceeded to calculate, for each bin, the probability of failure conditioned to Legal Form, next by Ateco Name and finally by Location.
Also in this case we proceeded to apply the binomial test. In those cases, the sample proportion is the number of failed firms in a specific subgroup (e.g., Location = 'Center' and Size = 'Small'), while the hypothesized proportion is the probability of failure rela-

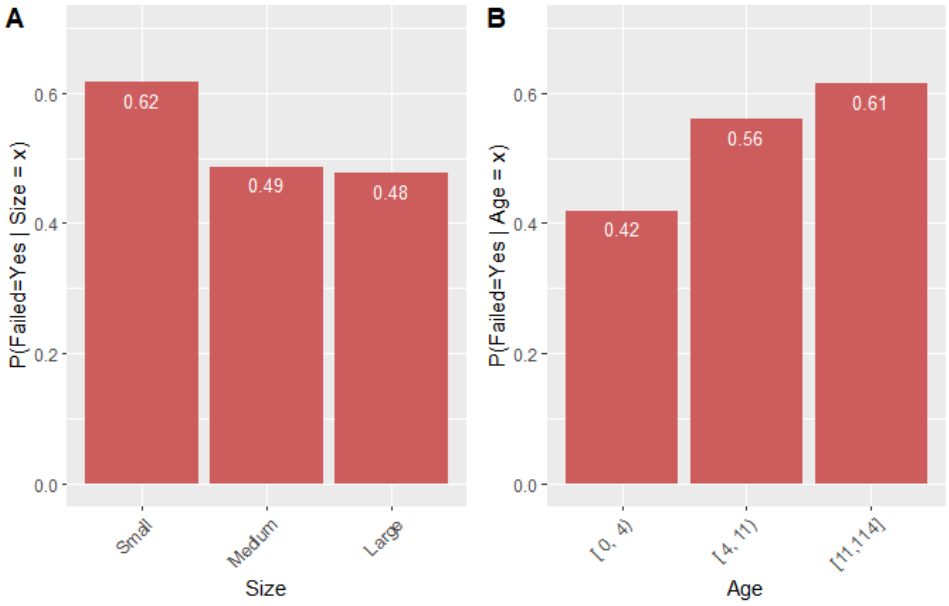

Figure 5: Conditional Probability of being failed over Age and Size

tive to the size of the firm (e.g., Size = 'Small'). The null hypothesis to be tested is then whether these two proportions are equivalent. The significance level was again corrected with Bonferroni for multiple binomial tests, and it is reported in each table.

Considering the probability conditional on Size and Location, we see in Table 5 that the null hypothesis is rejected in all cases except for the (Size = Small) and (Location = Center) subgroup. This tells us that for firm of every size type (small, medium, large) there is a statistically significant difference between the probability of default (PD) of a firm given its size type and the PD obtained when further conditioning on most location. This suggest that the firm's location is an influential predictive feature of PD.

The same phenomenon can be observed by conditioning on Legal Form. For all company size, the probability of failure is statistically different when conditioning on most companies' legal forms. For instance, we see that when conditioning on (Size = Small), the probability of default (P = 62%) is statistically different from all other PDs obtained when further conditioning on all companies' legal forms, except for Consortium companies for which we can not reject the hypothesis that the PD is equal to 62%.

When conditioning by Ateco Sector, we again see many sectors do provide a statistically difference. By looking at the non significant ones, we can see that for every firm size type (small, medium, large), there is no significant difference of the PD when further

SIZE			
	SMALL P = 0.62	MEDIUM P = 0.49	LARGE P = 0.476
Location	p-value (alpha = 0.01) (estimate)	p-value (alpha = 0.01) (estimate)	p-value (alpha = 0.01) (estimate)
Centro	ns	5.34e-12 (0.44)	3.8e-07 (0.44)
Isole	5.7e-4 (0.48)	8.09e-20 (0.38)	2.6e-20 (0.35)
Nord - Est	2.6e-20 (0.74)	5.62e-20 (0.60)	1.8e-20 (0.56)
Nord - Ovest	3.8e-20 (0.73)	4.34e-20 (0.59)	2.8e-20 (0.56)
Sud	6.43e-71 (0.51)	5.39e-39 (0.40)	1.5e-20 (0.39)

Table 5: Binomial Test. H0: the Location proportion of failed companies of a given size is equal to the conditional probability of failing on that given size

SIZE			
	SMALL P = 0.62	MEDIUM P = 0.49	LARGE P = 0.47
Legal Form	p-value (alpha = 0.007) (estimate)	p-value (alpha = 0.007) (estimate)	p-value (alpha = 0.007) (estimate)
Consortium	ns	ns	2.8-12 (0.33)
Other	2.90e-20 (0.17)	1.3e-20 (0.13)	5.3e-20 (0.10)
S.C.A.R.L.	1.6e-6 (0.57)	ns	5.3e-6 (0.52)
S.P.A.	1.0e-4 (0.9)	1.5e-5 (0.77)	7.9-e14 (0.66)
S.R.L.	1.06e-20 (0.71)	8.2e-20 (0.52)	ns
S.R.L. one-person‘	3.65e-20 (0.82)	5.9e-20 (0.61)	1.8e-6 (0.51)
S.R.L. simplified	0 (0.38)	2.4e-20 (0.247)	1.2e-15 (0.33)

Table 6: Binomial Test. H0: the Legal Form proportion of failed companies of a given Size is equal to the conditional probability of failing on that given size

conditioning on industry sector 'H - Trasporti', 'PB - Educazione & Miniere' and 'Q - Sanità'. This tells us that knowing that a firm belongs to any of those industry sector does not influence our initial PD estimate based on a firm's size.

SIZE			
	SMALL P = 0.62	MEDIUM P = 0.49	LARGE P = 0.47
Ateco Name	p-value (alpha = 0.004) (estimate)	p-value (alpha = 0.004) (estimate)	p-value (alpha = 0.004) (estimate)
A	7.4e-10 (0.48)	1.0e-13 (0.3)	6.6e-20 (0.25)
EL	2.4e-13 (0.69)	1.2e-15 (0.56)	4.3e-8 (0.43)
FN	9.3e-7 (0.58)	ns	ns
GM	4.4e-9 (0.64)	2.0e-3 (0.50)	5.9e-6 (0.50)
H	ns	ns	ns
I	4.2e-20 (0.50)	1.8e-20 (0.37)	1.5e-18 (0.35)
JC	7.8e-18 (0.68)	2.50e-12 (0.54)	5.3e-20 (0.56)
KD	9.9e-06 (0.69)	7.43e-7 (0.6)	ns
PB	ns	ns	ns
Q	ns	ns	ns
R	ns	4.35e-6 (0.4)	ns
S	1.6e-07 (0.51)	2.0e-4 (0.4)	ns

Table 7: Binomial Test. H0: the ateco sector proportion of failed companies of a given Size is equal to the conditional probability of failing on that given size.

Age

The same approach was used for analysing the PD when conditioning on a firm's Age. We discretized the variable Age, which was divided into three bins with equal frequency ('[0, 4)','[4, 11)','[11,114]'). We then calculated the conditional PD over Age (Figure 5 B) and performed an all pairs comparison (with the binomial test) of the proportions of Failed when conditioning on the various Age categories. The confidence level was set at 98% to account for the multiple tests and the alternative hypothesis was set to "less" for each test. A P-value of 2.2e-16 was obtained for all tests and we thus rejected the null hypothesis of equal proportions, concluding that the PD varies significantly with the age of a firm.

We investigated how the probability of failure conditional over Age changes by further conditioning on Legal Form, Location and Ateco Sector. We performed multiple binomial tests comparing the PD for each Age category to the various PDs obtained when further conditioning on the various categories of the other variables.

The test results for the conditional probabilities over Age and Legal Form (Table 14 in Appendix) show that in almost all subgroups there is a statistical difference with respect

to the respective PD of the Age group. A noticeable exception can be seen for S.P.A. companies where, for each Age group, we could not reject the hypothesis that the PD obtained was equal to the ones of their respective age group. Thus, for a firm of any age group, knowing that the firm is also S.P.A. does not significantly change our estimated PD.

When conditioning the PD over Age and Location (Table 15 in Appendix), we see that for all subgroups there is a statistically significant difference with the conditional probability of their respective Age value, without any exception. This tells us that when conditioning on a firms age, any additional knowledge about a firms' location results in a significant change in the estimated PD.
Finally when conditioning on Ateco Form (Table 16 in Appendix), we see that in almost all subgroups there is a statistically significant difference.
In particular we noticed that for firms who's Ateco sector is 'PB - Istruzione & Miniere' and 'H - Trasporti', we can not reject the null hypothesis for any age group. Meaning that for a firm of any age group, knowing that it operates in any of the previous sectors does not significantly change the PD estimate related to its age group.

3.4 Question D

Scoring and rating models for assessing risk of default

3.4.1 Data preprocessing for predictive tasks

We first removed the insignificant features for our predictive models such as "Company name" and "Tax code number".
As many of the features in our data set were repeated measures of financial indicators taken in consecutive years. We thus expected many metrics to be highly correlated. Our intuition was confirmed by a correlation analysis. All measurements at last available year - 2 were dropped due to they very high number of missing values. We further transformed the repeated measurements by keeping only the measurements at the last available year and the difference between the measurements at the last available year and the previous one (new column "Avg. trend") so as to provide information on the trend of the financial performance of firms.

We then subselected the years which we would later use for training (2010 until 2016) and test (2017). We chose these years as they provided a balanced mix between percentage of failed firms and number of total firms.
We then imputed missing values. Given the high amount of missing values, we tried to find the balance between dropping entire variables and removing the rows with missing values. Our goal was to obtain a balanced dataset with an appropriate large number of features whilst having enough rows for concept learning. We thus fined-tuned by hand a threshold which would provide us such a data set and proceeded to remove the variable who's missing values were above this value and then we removed the remaining missing rows.
The new data set has 141'000 rows and 51 variables.

The next step was feature selection. We first proceeded to drop predictors that were significantly independent from the dependent variable "Failed". We temporarily binned continuous distribution then used the G-Test of independence. One financial indicator was found to be independent and thus dropped.
We then controlled whether predictors were correlated between each other. From a correlation plot we saw a high risk of multicollinearity. We calculated the Variance Inflation Factor (VIF) for every variables and proceeded to remove variable one by one until all features had a VIF under 5.
The final feature selection step was performed with the backward STEP AIC algorithm. We repeatedly learned a linear model on our data and removed the variables which increased the overall Akaike information criterion (AIC). However, the marginal diminishment in AIC were so small that we decided not to remove any variable based on the STEP AIC.

The final step consisted in temporally splitting the data set into training and test data sets. Our training set ranged from year 2012 to 2014 and had 70'490 rows. The test set only contained firms from 2015 and had 24'000 rows.
Particular attention was given to covariate shift between the training and test samples. The probability of failure in each set was respectively 62.4% and 52.6%. We thus re-balanced the training set, by randomly removing failed firms, to meet the same balancing of the test set. The final training set had 56'610 rows.

3.4.2 Parametric models

Logistic regression

We fitted a logistic regression model on the training sample. We visually inspected the residual error to assert whether the normality of the error assumptions for the parametric model were met. As seen in Figure 6, the binned residual are centered around zero and are generally contained in the normal variance.

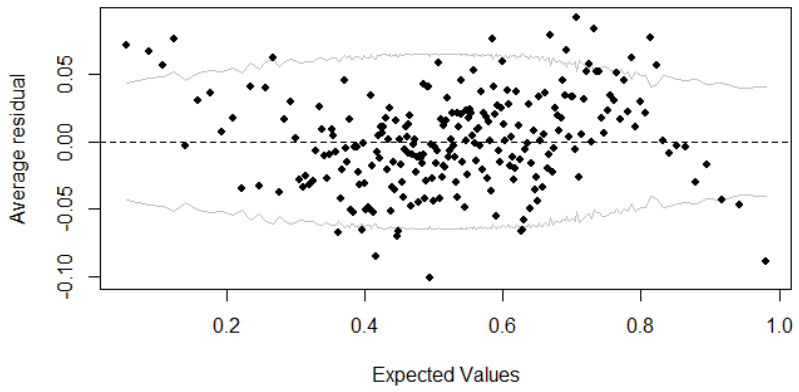

Figure 6: Binned residual error of non-penalized logistic regression.

We then performed a post-hoc test to check the model's coefficients along with their relative p-values for assessing their significance. We saw that with a 95% confidence, for 18 predictive features out of 52 we could not reject the null hypothesis that the coefficients were equal to zero thus labelling them as not statistically significant for our model.
In table 16 of the Appendix, we reported the estimate of the coefficient of the statistically significant predictive features, their standard error and p-value. The H0 is that the

coefficient associated with the feature is zero, meaning that the variable has no correlation with the dependent variable. Note that p values has been adjusted to take into consideration multiple tests. For example, the firms which are located in Nord-Est have a coefficient β of 0.558. Meaning that, ceteris paribus, they have $e^{0.558} = 1.747$ times the odds of the firms which are located elsewhere to fail. Particularly surprising it the fact that firms which operate in Ateco sector "JC - Comunicazioni and Manifattura" have as much as 2.75 times the odds of firms which have another Ateco sectors to fail. Instead, for firms which have Legal Status equal to Other, we can observe that that beta is a negative number, and so we can assert that, ceteris paribus, those firms have lower odds of failure compared to other firms. To be precise, firms with Legal form equal to Other have 0.097 (approximately a tenth) times the odds of failing compared to firms with a different legal form.
The intercept is equal to -2.57 and it means that, assuming a value of 0 for all the predictors in the model, the probability of a firm to fail is 0.07 (7%).

Penalized Logistic regression
Given the numerous predictive features that were not significant in the logistic regression model, we though to investigate the effect of a Lasso penalized logistic regression.
The lambda parameter for the lasso regularization was selected through cross-validation by selecting the largest value of lambda such that error is within 1 standard error of the minimum. This was made to enhance model simplicity. By applying the Lasso logistic regression, 17 predictive features got their coefficient regularized to zero.
We noticed that 13 out the 18 variables that were statistically insignificant in the non-penalized regression have been regularized to exactly zero by the lasso logistic regression.

Assessment of parametric models' scoring performance
The estimate performance of the logistic regression on the test set are AUC ROC (Area under ROC curve) 0.679 and Bin ECE (binary expected calibration error) 0.014.
Those of the performance of the lasso logistic regression were AUC ROC 0.678 and Bin ECE 0.015.
In Figure 7, we can see the ROC curves (A) and calibration plots (B) for both the parametric models.

The Roc Curve is the scatter plot of True positive rate (TPR) against False Positive rate (FPR) for varying threshold level. In our case, we can see that, for instance, to reach a TPR of 75%, we have to account a FPR of around 50%. In order words, to correctly classify 75% of failed firms, around 50% of non-failed firms will be classified as failed.
The AUC ROC represents the area under the curve and it is the degree of separability, and it shows how good is the model in distinguishing between active and failed companies. We can see that both curves of the parametric classifiers almost completely overlap. Thus showing that both models have the same degree of separability, which was expected by

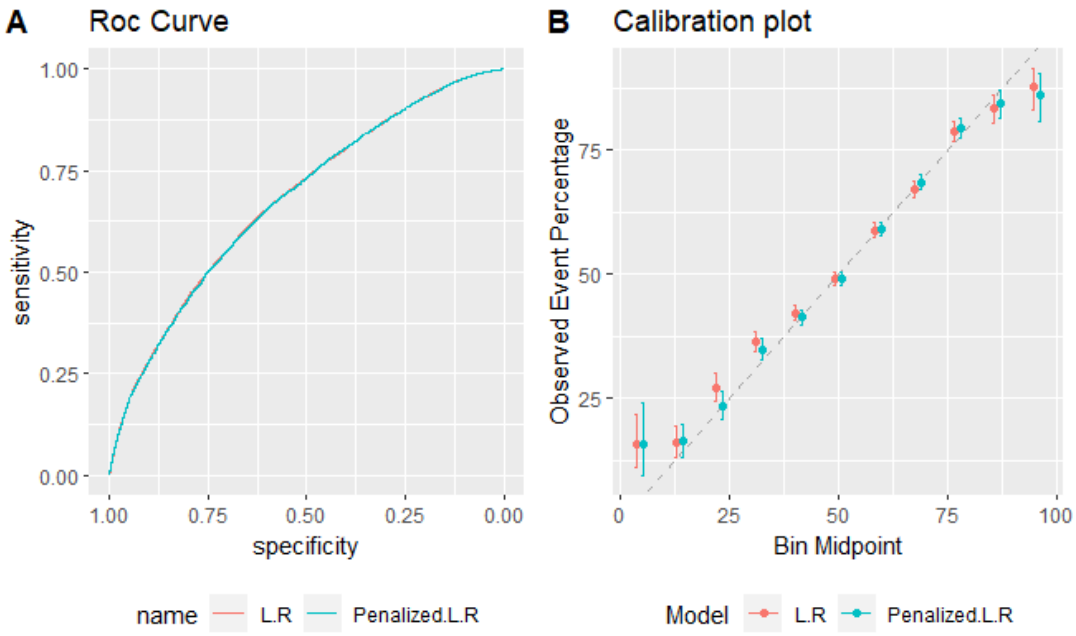

Figure 7: Roc Curve and Calibration Curve for Logistic Regression and Penalized Logistic Regression.

the nearly identical estimated AUC ROC. From the ROC curve we also see that the classifier do behave better than a random classifier.
The calibration plots compares the model's binned predicted probabilities against the observed ones. We can observe that both models are overall well calibrated, staying around the dotted line at 45 degrees which indicates perfect calibration where predicted and observed probabilities are equal. More precisely, we see that when the predicted probabilities are between 30% and 70% they lay over the ideal line. Instead, when the predicted probabilities are under 10% and above 80, we notice a slight diversion from the ideal line. For scores under 10%, both models underestimate the probabilities of failed companies. Instead, when scoring over 80%, the models tend to underestimate the probability of failing. This said, looking at the confidence interval of the binned probabilities, obtained through the binomial test, we observe that bins who's midpoints is in these extreme regions (i.e. where our classifiers predict a score of under 10% and above 80%) have a greater confidence interval and thus a greater uncertainty. This is due to the fact there is less data available in those bins because few firms receive extreme scores from the models. Therefore, although the parametric models tend to underestimate the true probability of failure for low scores, one should also recognize a higher uncertainty associated the true risk of failure for those bins.

3.4.3 Non-parametric models

Random Forest and AdaBoost

The first non-parametric model we decided to use was the Random Forest, a tree-based ensemble model that relies on the bagging ensemble method. For the Random Forest implementation, we performed a cross-validated grid-search to find the best parameters for the percentage of features to be randomly sampled as candidates at each split and the total number of trees of the ensemble. The final parameters of the model were respectively 30% and 50 trees. The model was then retrained on the entire training set.
An identical procedure was done for Adaboost, another tree-based ensemble model that instead relies on the boosting ensemble method. The parameters selected through grid-search were the number of trees and the maximum depth of each tree. The final parameters selected were 150 trees and a maximum depth of 4.

Assessment of non-parametric models' scoring performance

The final models were retrained on the entire training set and the following scoring performance were obtained on the test set.
Random Forest achieved a AUC ROC 0.718 and binECE 0.017, while AdaBoost achieved a AUC ROC 0.703 and binECE 0.025.

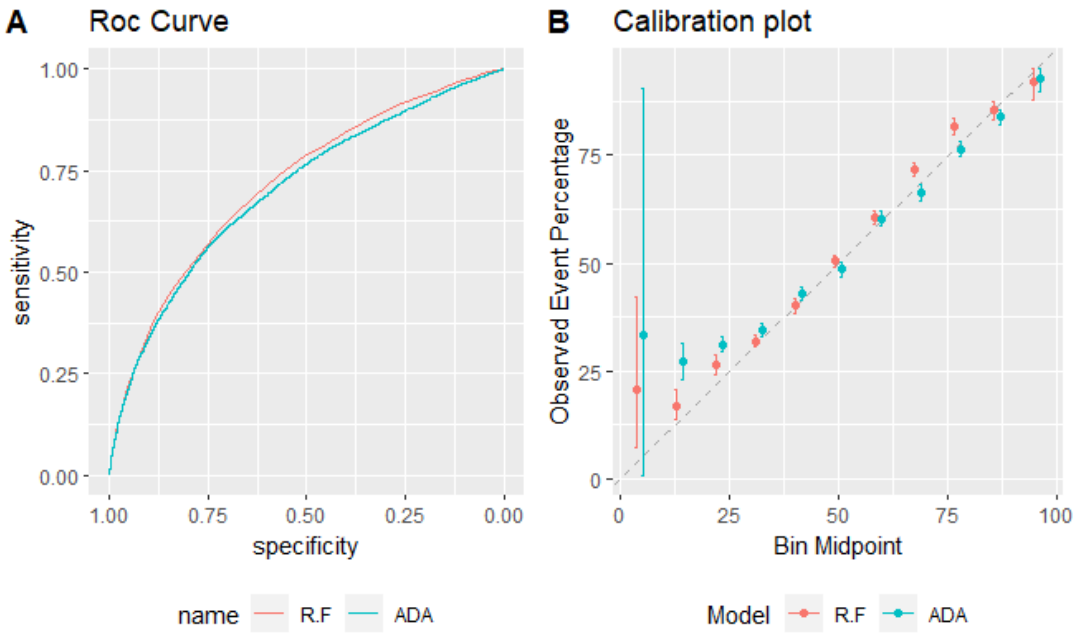

Figure 8: Roc Curve and Calibration Curve for Random Forest and AdaBoost.

In Figure 8, we can see the ROC curves and calibration plots for both models. We see a slight overlap of the Roc curves of both non-parametric models. The calibration plot also reveal a similar behaviour of the models. Overall we can see that the models are well-calibrated with predicted scores above 25% closely following the observed failure probability. For scoring values under 25% we see a strong tendency to underestimate the true risk of failure along with an almost maximum uncertainty associated to very low scoring predictions due to the fact that both non-parametric models have extremely few scores under 5%.

3.4.4 Models comparison

Once we assessed the performance of each classifier on the test set, we thought necessary to check whether there was a statistically significant difference among the classifier's scoring performance.
In order to obtain a repeated sample of performances for each estimator, we ran a repeated 10 fold cross-validation on the entire training set returning the AUC obtained on each fold. We ensured to use the same seed while we repeated the procedure for each classifier and we thus obtained a sample of 50 repeated AUC measurements for each classifier on the same folds.
We then used the Shapiro and Barlett test to control the hypothesis of normality of equal variance of the samples. Both tests provided high p-values, we therefore did not reject the former hypotheses. Given that our samples consisted of repeated measurements and that they satisfied the assumptions for a parametric test, we proceeded to do the comparison with a within-subject ANOVA as omnibus test and multiple paired t-tests with Bonferroni correction as post-hoc test as recommended by several texts (Keppel & Wickens, 2004; Maxwell & Delaney, 2004).
The p-value for the ANOVA F-test was equal to 2.31e-98, thus conveying that at least one sample had a significantly different mean. The p-values of the all-pairs post-hoc tests were all smaller than 0.05 except for test between non-penalized and penalized logistic regression where the p-value equaled 0.30 and we could not not reject the hypothesis that the models had the same average performance. All other tests' p-values were significantly small to reject the null hypothesis than any other sample had equal mean.
In Figure 9, we plotted the boxplot of the model's performance along with the statistically significant differences resulting form the all-pairs test. As anticipated, all pairs are statistically different except between the parametric models. The statistical difference between the parametric and non-parametric is visually straightforward and in fact all p-values associated to test between any parametric and non-parametric is close to zero. From the p-values of the post-hoc test, we also gain knowledge about the statistical significant difference in performance between the non-parametric models where the Bonferroni adjusted t-test between the Random Forest and AdaBoost classifiers had a P-value of 4.06e-9.

Concluding on the model comparison, we used AUC ROC as an estimator of the AUC[3] of each model, thus assessing their separability capabilities between failed and active firms. We asserted that there is sufficient evidence to reject the null hypothesis that all models had equal separability capabilities. Furthermore, we also rejected the hypothesis that the RandomForest and AdaBoost models had equal separability capability whilst we could not reject the hypothesis that non-penalized and penalized regressors had the same separability capability.

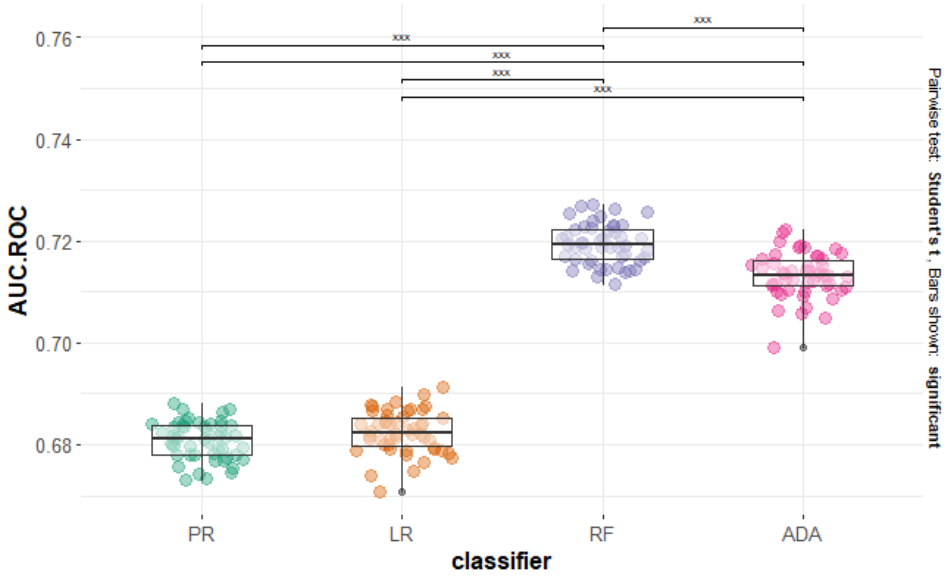

Figure 9: Model scoring comparison based with within-subject ANOVA and multiple paired t-tests with Bonferroni correction.

3.4.5 Rating Models

The final step in solving this task is the building of a rating model. For each classifier, the probability of default (PD) is mapped into a bucket for a total of nine buckets ranging from letter 'A' to letter 'I', each defined for an increasing range of probabilities of default. The upper limit of each bucket is reported is the column 'Threshold' in Table 8.
For each bucket, we assessed the degree to which the probabilities of default established by the various classifiers matched the observed realized defaults rate.
Referring to the strategy outlined in the paper Corporate Default Forecasting With Machine Learning, we performed multiple binomial test corrected with Bonferroni, one

[3]AUC is the estimates of the probability over the true population that being presented a failed and a non-failed firm, the score (assessing probability of failure) given by the model is able to actually be larger for the failed firm than the active.

for each of the defined buckets, and tested how compatible the probabilities of defaults are with the realized defaults rate. The null hypothesis was that the probabilities of defaults were below or equal to the thresholds. The alternative hypothesis was that probability of defaults are greater than the thresholds.
In Table 8 we reported the results of the binomial tests, each cell indicates the percentages of realized default rate that are within the bucket. Based on the p-value of the test, we assigned the color green to cells to indicate that realized defaults rate are below or equal to the expected threshold (when p-value is greater than 0.20); the color yellow indicates that there is a relative discrepancy between expected and realized defaults (p-value below 0.1) and the red one even a strong one (p.value below 0.01). The results showed that for firms which have been classified with rating 'D', 'E', 'F', 'G', 'H', 'I', the estimated probabilities of default match (or are lower) the realized default rate for all classifiers. Instead, for class 'A', where the models' estimates of default probability are lower, there is a statistically significant discrepancy between the models' default scores and the realized default rate for all models. For instance, firms that were classified as 'A' by the parametric models (when the model predicted a failure score between 0% and 10%), had an observed default rate of 17% deemed significantly different (by the binomial test) from the 10% threshold of bucket 'A'. This problem is even more evident for the AdaBoost model that predicted had 50% realized default rate for bucket 'A' (the higher p-value shown by the yellow color is due to the very few firms in bucket 'A') and also underestimates the true default risk for bucket B and C. This is mitigated by the other models, which tend to be more precise when making predictions in bucket 'B' and 'C'.
A probable cause of significant discrepancy observed for the models for bucket 'A' is the very few firms with an associated score of less than 10%. The low number of data increase the uncertainty in the true default rate. The results of Table 8 are in fact related the ones observed in the calibration plot of the models (Figure 7 and Figure 8). As seen in Figure 7 and Figure 8, the models underestimate the true probability of default when scoring under 10% and parametric models have a much narrower confidence interval than non-parametric for such scores. This is seen in Table 8 where parametric models have a much lower p-values (color red) and this a stronger statistical difference the threshold of the related bucket.

Bucket	Threshold	Non-penalized Logistic Regression	Penalized Logistic Regression	RandomForest	AdaBoost
A	10%	17%	17%	19%	50%
B	20%	17%	15%	20%	27%
C	30%	30%	27%	28%	32%
D	40%	39%	37%	35%	36%
E	50%	45%	44%	45%	46%
F	60%	55%	55%	58%	55%
G	80%	69%	70%	74%	68%
H	90%	82%	85%	86%	82%
I	100%	80%	85%	90%	91%

Table 8: Binomial Test. H0: true probability of defaults is equal to the thresholds, for each bucket and each model. H1: greater. The threshold column indicates the upper limit of the our bucket interval. For instance, a firm is classified as D if its default probability is between 30% and 40%. The percentages in the colored cells represent the realized default rate for each bucket. Cell colored in green indicates a p-value greater than 20%, yellow cells a p-value between 1% 20%, red ones a p-value less than 1%.

3.5 Question E

Extend/investigate the scoring models

In this final task, we decided to investigate whether applying selective classification to our models had the same effect among all the classifiers.
When making a selective binary classification, we use a so-called reject option where the classifier decides to abstain when the score predicted is too close to 50% (i.e. maximum uncertainty). By increasing the rejection zone, our classifier will only classify instances where he is increasingly certain, thus increasing its classification performance.
As we can see in the risk-coverage curve plotted (Figure 10 A) for our logistic regression model, the increase in performance (decrease in risk), comes with a proportional loss in coverage. For instance, we can see that allowing a coverage of only 50% would decrease the misclassification error of our regression model by 10 percentage points.

In order to measure and compare the effect of applying selective classification among all our models, we chose accuracy as our classification performance measure and set a minimum support threshold of 50% for all classifiers.
The statistical question we would thus answer is: "setting a coverage of 50%, do all models obtain the same marginal gain in accuracy when applying selective classification?"

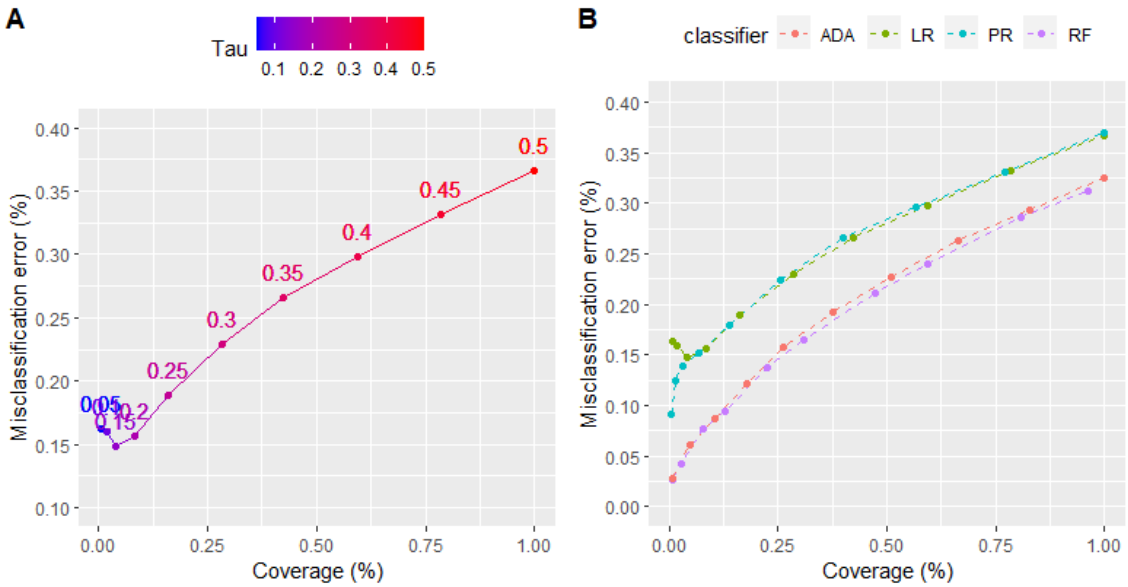

Figure 10: Risk-coverage curve for logistic regression model assessed on test set

To answer this question, we measured, through cross-validation, the accuracies of each classifier when coverage was 100% (Tau set to 0.5) and when coverage was 50%.
To measure the latter, we created a custom function that would find, for each model and for each fold, the highest value of Tau which would result in a 50% coverage. For instance, the average of the values of Tau used in each fold for the logistic regression selective classification problem was around 0.375 which is in line with what the reader can observe in Figure 10 A.
In each fold, we calculated the marginal gain in accuracy by subtracting the accuracy obtained with 100% coverage from the accuracy obtained with 50% coverage.
By repeating the same process for each classifier on the same folds (by setting the same seed), we eventually obtained a data set of paired measurements which represented the marginal gains in accuracy for each classifier on the same fold.

We could now proceed to compare the marginal gains. After asserting that the data satisfied the assumption of normality and equal variance through the Shapiro and Barlett's test, we proceeded with a within-subject ANOVA followed by a multiple paired t-tests with Bonferroni correction.
The F-test of the within-subject ANOVA had a P-value of 1.05e-4. We thus reject the hypothesis that all the samples have equal mean.
The pair-wise paired t-tests revealed that there is a statistically significant difference between the marginal gain of parametric and non-parametric models; however there is

no statistically significant difference within the two groups (Figure 11). The paired t-test between parametric models returned a p-value of 0.21 and the paired t-test between the non-parametric models returned a value of 1. All other comparisons had adjusted p-value low enough to reject the hypothesis of equality.

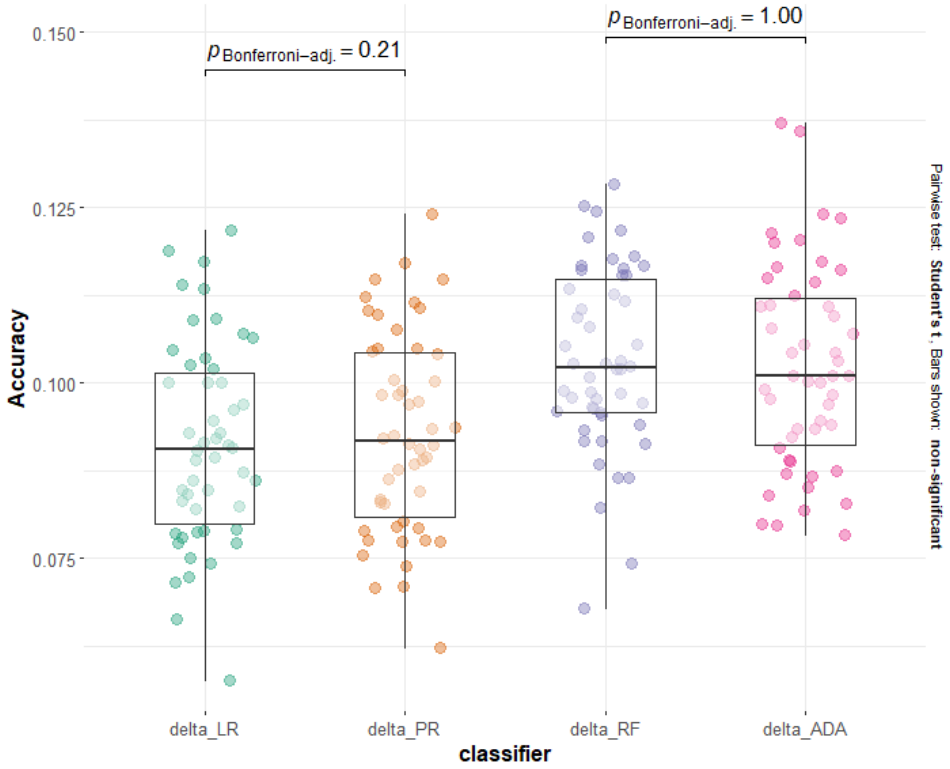

Figure 11: Boxplot of marginal gain in classification performance of each classifier when applying selective classification. Top bars are shown only for NON significant comparisons.

The significant difference between the marginal gain in classification accuracy between parametric and non-parametric models identified by the post-hoc test does indeed coincide with what the reader can observe in Figure 10 B. Taking a close look, we can see that the gap between parametric and non-parametric models' misclassification error at 100% coverage is slightly less than 0.05. Whilst at 50% coverage, the gap increases to about 0.07, thus showing that the non-parametric model's have a higher marginal gain in accuracy (i.e. decrease in misclassification error) when reducing coverage from 100% to 50%.

We conclude by answering the question we asked our-self's earlier by stating that there is enough statistical evidence to reject the hypothesis that all our classifiers obtain the same marginal gain in classification performance when applying selective classification

with a coverage of 50%. The increase in accuracy for both non-parametric models was statistically different compared to the parametric models.
After assessing in Question D that the average scoring performance of the parametric models did not match the scoring performance of non-parametric models, we have now assessed they their gain in classification performance, when applying a binary selective classification, does not match the gain obtained by non-parametric models.

References

[1] Mirko Moscatelli and Simone Narizzano and Fabio Parlapiano and Gianluca Viggiano, 2019. *Corporate default forecasting with machine learning* 1256, Bank of Italy, Economic Research and International Relations Area.

[2] Keppel, G.,& Wickens, T.D. (2004). *Design and analysis: A researcher's handbook (4th ed.).* Upper Saddle River, NJ: Prentice Hall.

[3] Maxwell, S.E ., & Delaney, H.D . (2004).*Designing experiments and analyzing data: A model comparison perspective (2nd ed.).* Mahwah, NJ: Lawrence

4 Appendix

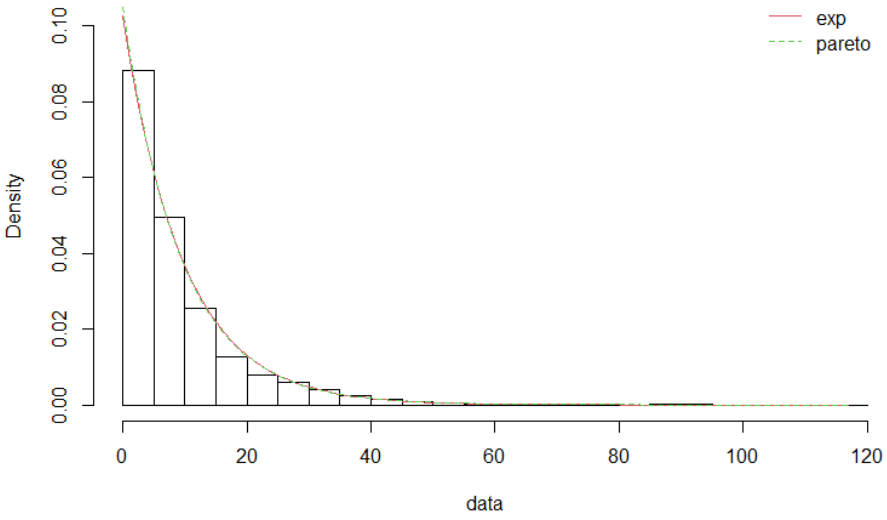

Figure 12: Pareto and Exponensial Fitting on Age Feature

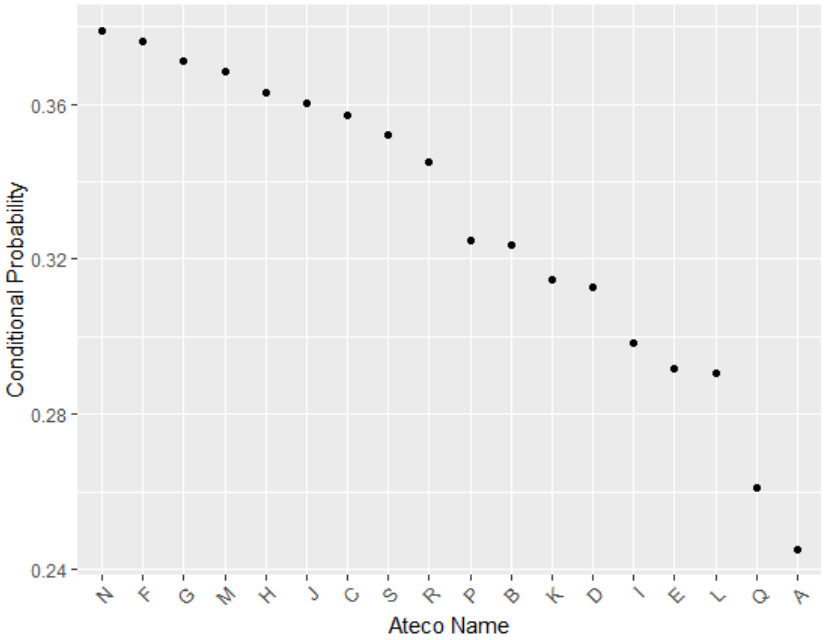

Figure 13: Probability of Failure conditioned on Ateco Name

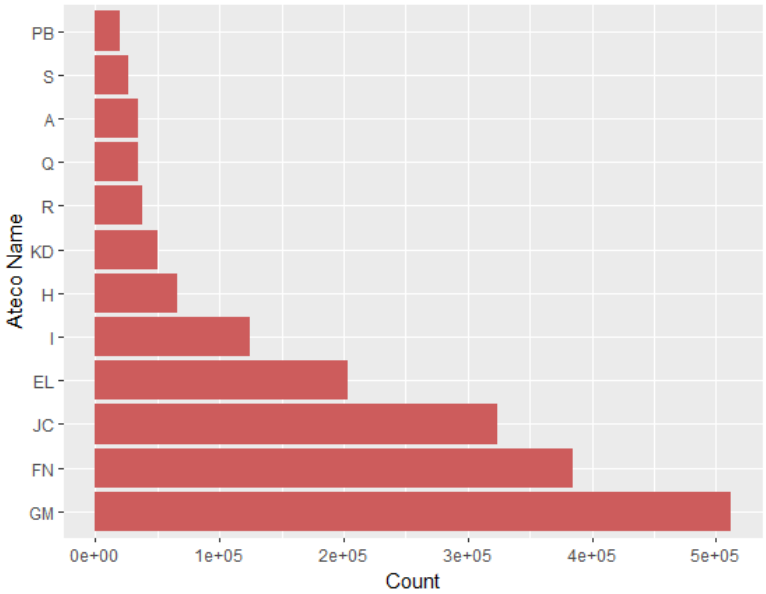

Figure 14: Bar Plot of Ateco Name.

ATECO NAME
A - Agricoltura
EL - Acque&Immobiliari
FN - Costruzioni&Noleggio
GM- Ingrosso&Scientifico
H - Trasporti
I - Ristorazione
JC - Comunicazioni&Manifatture
KD - Assicurazione&Energia
PB - Istruzione&Miniere
Q - Sanità
R - Arte&Sparti
S - Altri Servizi

Table 9: Ateco Name.

AGE				
	Active	Failed	Difference	Statistical significance
Legal Form	Mean	Mean	99.3% CI	p-value (alpha = 0.007)
Consortium	20.05	12.73	(5.34 , 9.31)	3.8e-20
Other	15.08	14.03	(-1.55 , 3.64)	ns
S.C.A.R.L.	9.48	11.06	(-2.35 , -0.80)	4.1e-8
S.P.A.	22.12	25.15	(-7.48 , 1.42)	ns
S.R.L.	9.20	11.80	(-2.84 , -2.34)	5.6e-20
S.R.L. one-person	10.40	12.03	(-2.24 , -1.08)	1.9e-14
S.R.L. simplified	1.01	1.26	(-0.31 , -0.19)	6.1e-20

Table 10: T-test for a specific Legal Form. H0: true mean Age difference between active and failed companies is equal to zero.

AGE				
	Active	Failed	Difference	Statistical Significance
Ateco Name	Mean	Mean	99.6% CI	p-values (alpha = 0.004)
A	10.52277	13.19	(-4.66 , -0.68)	1.0e-4
EL	15.14654	16.32	(-2.03 , -0.34)	6.00e-05
FN	8.054809	11.21	(-3.58 , -2.74)	3.25e-20
GM	6.735685	9.72	(-3.35 , -2.6)	1.5e-20
H	6.572881	9.220	(-3.61 , -1.68)	4.2e-20
I	5.196577	6.36	(-1.71 , -0.62)	8.7e-10
JC	8.73	12.53	(-4.43 , -3.18)	4.4e-10
KD	7.994135	9.60	(-3.04 , -0.18)	1.0e-3
PB	9.118557	9.969	(-2.83 , 1.13)	ns
Q	8.197962	8.66	(-1.91 , 0.98)	ns
R	6.798287	8.26	(-2.60 , -0.32)	2.4e-3
S	5.317439	7.98	(-3.90 , -1.41)	9.4e-10

Table 11: T-test for a specific Ateco Sector. H0: true mean Age difference between active and failed companies is equal to zero.

AGE			
	Shift (2014 - 2016)		Statistical significance
Legal Form	estimate	99.1% CI	p-value (alpha = 0.008)
ALL*	1.8e-5	(4.6e-5,1.3e-5)	4.44e-14
Consortium	-1.99	(-2.99, -0.99)	9.3e-6
S.C.A.R.L.	6.7e-5	(-5.7e-5, -9.9e-1)	ns
S.P.A.	-3.0	(-6.0,-0.99)	1.0e-3
S.R.L.	-0.99	(-0.98 , -0.99)	9.4e-20
S.R.L. one-person	-2.0	(-0.04 , 0.16)	1.9e-20

Table 12: Wilcoxon rank-sum test. H0: Age distributions of 2014 and 2016 considering a specific legal form are related by a shift in location, that is the difference of the two expectations, and this shift is zero.

AGE			
	Shift (2014 - 2016)		Statistical significance
Location	estimate	99.2% CI	p-value (alpha = 0.008)
ALL*	1.8e-5	(0.08 , 0.16)	4.4e-14
Centro	-2.2e-5	(-3.4e-5, 1.0e-5)	ns
Isole	-3.2e-6	(-9.9e-1, 5.7e-5)	ns
Nord - Est	-7.0e-6	(-9.9e-1, -5.2e-5)	1.0e-3
Nord - Ovest	-0.99	(-9.9e-1, -3.4e-5)	1.6e-10
Sud	2.5e-5	(-6.1e-5 , 4.5e-5)	ns

Table 13: Wilcoxon rank-sum test. H0: Age distributions of 2014 and 2016 considering a specific location are related by a shift in location, that is the difference of the two expectations, and this shift is zero.

AGE			
	[0, 4) P = 0.42	[4, 11) P = 0.56	[11,114] P = 0.61
Legal Form	p-value (alpha = 0.007) (estimate)	p-value (alpha = 0.007) (estimate)	p-value (alpha = 0.007) (estimate)
Consortium	1.6e-3 (0.51)	ns	1.6e-12 (0.41)
Other	2.6e-20 (0.12)	3.6e-20 (0.14)	2.2e-20 (0.12)
S.C.A.R.L.	5.4e-20 (0.49)	8.0e-10 (0.55)	ns
S.P.A.	ns	ns	ns
S.R.L.	3.5e-20 (0.46)	4.8e-20 (0.55)	3.0e-9 (0.63)
S.R.L. one-person	2.0e-4 (0.54)	ns	1.2e-10 (0.67)
S.R.L. simplified	4.9e-20 (0.33)	1.6e-17 (0.43)	ns

Table 14: Binomial Test. H0: the Legal Form proportion of failed companies of a given age is equal to the conditional probability of failing given that age.

AGE			
	[0, 4) P = 0.42	[4, 11) P = 0.56	[11,114] P = 0.61
Location	p-value (alpha = 0.01) (estimate)	p-value (alpha = 0.01) (estimate)	p-value (alpha = 0.01) (estimate)
Centro	9.0e-3(0.40)	1.4e-5 (0.53)	1.4e-9 (0.57)
Isole	1.8e-20 (0.32)	1.7e-20 (0.44)	4.8e-20 (0.49)
Nord-Est	1.4e-20 (0.50)	1.18e-20 (0.67)	1.4e-20 (0.71)
Nord-Ovest	2.5e-20 (0.51)	1.12e-20 (0.66)	4.6e-20 (0.68)
Sud	1.4e-20 (0.35)	2.4e-20 (0.47)	2.3e-20 (0.53)

Table 15: Binomial Test. H0: the Location proportion of failed companies of a given age is equal to the conditional probability of failing on that given age.

AGE			
	[0, 4) P = 0.42	[4, 11) P = 0.56	[11,114] P = 0.61
Ateco Name	p-value (alpha = 0.004)	p-value (alpha = 0.004)	p-value (alpha = 0.004)
A	7.4e-10 (0.29)	1.8e-8 (0.30)	5.3e-20 (0.42)
EL	2.4e-13 (0.39)	ns	3.6e-14 (0.56)
FN	9.3e-7 (0.37)	1.2e-12 (0.53)	ns
GM	4.4e-9 (0.44)	8.6e-9 (0.59)	5.07e-16 (0.66)
H	ns	ns	ns
I	4.2e-20 (0.36)	9.8e-12 (0.49)	1.6e-19 (0.48)
JC	7.8e-18 (0.48)	4.7e-16 (0.61)	1.46e-20 (0.68)
KD	9.9e-06 (0.51)	5.12e-6 (0.62)	ns
PB	ns	ns	ns
Q	ns	ns	1.0e-3 (0.53)
R	ns	ns	8.18e-5 (0.52)
S	1.6e-07 (0.36)	ns	ns

Table 16: Binomial Test. H0: the Ateco Sector proportion of failed companies of a given age is equal to the conditional probability of failing on that given age.

Features	Coefficients	SE	P.values
(Intercept)	-2.569	0.156	0
LocationIsole	-0.213	0.033	0
LocationNord-est	0.558	0.029	0
LocationNord-ovest	0.564	0.025	0
Current.liabilities.Tot.ass..Last.avail..yr	0.721	0.05	0
Current.ratioLast.avail..yr	0.08	0.016	0
EBITDA.Vendite.Last.avail..yr	-0.002	0	0
Legal.formOther	-2.231	0.092	0
Legal.formS.C.A.R.L.	-0.319	0.078	0.002
Legal.formS.R.L. simplified	-0.801	0.115	0
Liquidity.ratioLast.avail..yr	0.124	0.017	0
Number.of.employeesLast.avail..yr	0.004	0.001	0
Return.on.asset..ROA..Last.avail..yr	-0.013	0.001	0
Solvency.ratio.....Last.avail..yr	-0.004	0.001	0
Total.assets.turnover..times.Last.avail..yr	0.226	0.013	0
Age	0.016	0.001	0
ATECO.NAMEEL	0.606	0.085	0
ATECO.NAMEFN	0.66	0.079	0
ATECO.NAMEGM	0.903	0.079	0
ATECO.NAMEH	0.639	0.088	0
ATECO.NAMEI	0.413	0.085	0
ATECO.NAMEJC	1.013	0.081	0
ATECO.NAMEKD	0.981	0.108	0
ATECO.NAMEPB	0.776	0.12	0
ATECO.NAMEQ	0.9	0.111	0
ATECO.NAMER	0.543	0.099	0
ATECO.NAMES	0.81	0.104	0
Size	0.06	0.008	0
Current.liabilities.Tot.ass...Avg.trend..yr	0.28	0.07	0.003
Liquidity.ratio.Avg.trend..yr	0.084	0.024	0.019
Profit..loss.th.EUR.Avg.trend..yr	0	0	0.036
Return.on.asset..ROA...Avg.trend..yr	0.003	0	0
Total.assets.turnover..times..Avg.trend..yr	-0.074	0.014	0
Total.assetsth.EUR.Avg.trend..yr	0	0	0.002

Table 17: Logistic Regression Model. H0: beta coefficients are equal to zero.