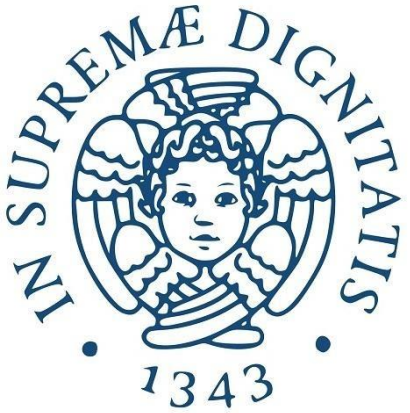

Università di Pisa

Statistical methods for Data Science project

Analisi del dataset AIDA

Un approfondimento sullo status legale delle imprese italiane

---

A cura di

Luca Palla

Sommario

1. Presentazione e preparazione dei dati

Il Dataset preso in esamine al fine dello svolgimento dell'analisi, è l'Aida. Tale dataset contiene al suo interno i dati di 1894412 aziende con corrispondenti 80 variabili. I dati raccolti si riferiscono agli ultimi 3 esercizi in cui l'azienda ha svolto rendicontazione. Le caratteristiche riportate spaziano in vari settori, da quello economico, a quello giuridico, da quello geografico a quello temporale. Le S.r.l., nelle loro varie forme, costituiscono la gran parte delle aziende all'interno della raccolta, rappresentando quasi il 90% del totale dei dati.

Al fine dell'analisi, è stato necessario intraprendere un'azione di selezione delle *features* con lo scopo di rendere la stessa meno dispersiva. Le colonne selezionate riguardano le seguenti informazioni: *Ateco 2007, Cash Flow, Current liabilities/Total assets, CurrentRatio, Interest/Turnover, Last Accounting Closing Date, Legal form, Legal Status, Leverage, NWC, Profit, Registered office address - region, ROA, ROE, Solvency Ratio, Tax code, numero di dipendenti, Ebitda/vendite, Total asset, Ebitda.* Tale scelta è stata fatta tenendo in considerazione sia il numero di missing values, sia le analisi già presenti in letteratura [1].
È stato deciso di implementare la variabile "**Status**" come target, andando a sostituire Legal Status. La struttura è binaria e i livelli ad essa attribuita sono "Active" o "Failed". All'interno della categoria *Failed* vi sono state inserite tutte le aziende che, nell'ultimo anno di rendicontazione, versavano in uno stato di: "Bankruptcy", "In liquidation", "Dissolved (liquidation)","Dissolved" , "Dissolved (merger)", "Dissolved (bankruptcy)". Di conseguenza, nella categoria *Active* sono state incluse "Active", "Active (default of payments)", "Active (Receivership)". Con tale suddivisione, ritroviamo 1190592 aziende *Active* e 703820 *Failed.*

Ulteriori variabili sono poi state create estrapolando informazioni dal dataset o raggruppando alcuni suoi valori:

- "**DimensioneAzienda**" divide il dataset fra "Micro" (dipendenti ≤ 10 e attivo ≤ 2 milioni), "Small" (10 ≤ dipendenti ≤ 50 dipendenti, 2 milioni ≤ attivo ≤ 10 milioni) e "Medium – Large" (dipendenti > 50 e attivo >10 milioni). In totale le aziende Micro rappresentano la gran parte del dataset, rispecchiando l'architettura aziendale tipica del territorio italiano.

- "**Età**" è costruita andando a prendere in considerazione "Last Accounting Closing Date" ed "Incorporation Year". Una volta creata, è stato notato che alcune aziende avevano un'età negativa, andando quindi contro i principi logici della variabile stessa. Per questo motivo, a tali osservazioni è stato attribuito il valore "NA'

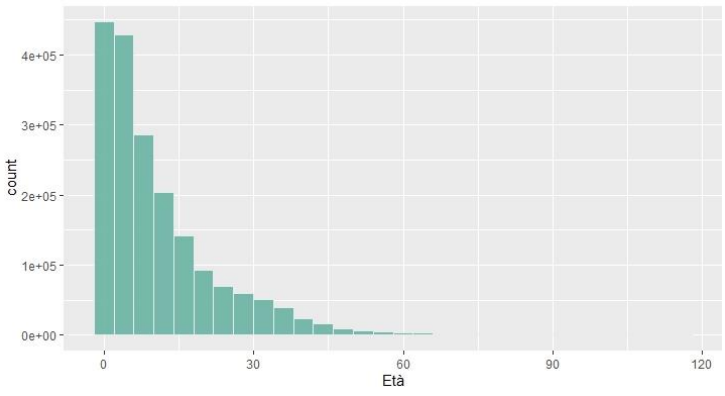

Figura 1.1 Istogramma Età

È stato poi deciso di creare differenti dataset: uno contenente solo i dati dell'ultimo anno di rendicontazione per ogni azienda e uno contenente solo l'anno precedente all'ultimo anno di rendicontazione (utile poi per la costruzione del classificatore finale).

1.1 Trattamento missing values e outliers

Prima di iniziare con le analisi, è stato però necessario trattare i missing values e gli outliers, in quanto molto presenti all'interno dei dataset.

Al fine di trattare i valori mancanti, ci siamo prima posti un altro interrogativo, e cioè se i nostri dati potessero seguire o meno l'assunzione di essere MCAR (missing completely at random) oppure no. La condizione MCAR è molto gradita in quanto sennò si rischia di inficiare la sostituzione dei missing. Difatti, per poter essere definiti MCAR, le distribuzioni dei due boxplot rappresentanti la distribuzione senza missing (azzurri), e quella con i missing (rossi), devono essere simili. Per quanto analizzato, tale condizione è plausibile all'interno del dataset. Nelle figure sottostanti riportiamo alcune delle analisi effettuate tramite il marginplot e il pbox della libreria **VIM**.

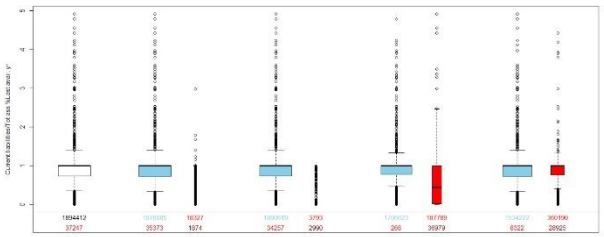

Figura 1.2 (a) Comparazione di più variabili

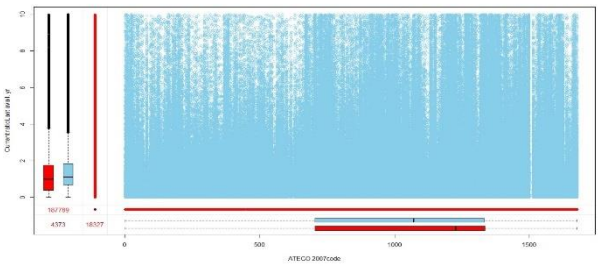

Figura 1.2 (b) comparazione Ateco e CurrentRatio

Il pacchetto utilizzato per la sostituzione dei missing values è "**mice**". Tale pacchetto permette di selezionare fra molteplici metodi come sostituire i valori mancanti. La metodologia utilizzata in tale analisi è il "**pmm**" (Predictive mean matching). I parametri impostati portano dunque alla costruzione di 3 dataset ognuno dei quali viene creato dopo 5 iterazioni.

Per poter valutare la bontà o meno della sostituzione sono stati utilizzate sia la comparazione della densità fra valori osservati e valori sostituiti e quella fra la distribuzione prima e dopo la sostituzione.

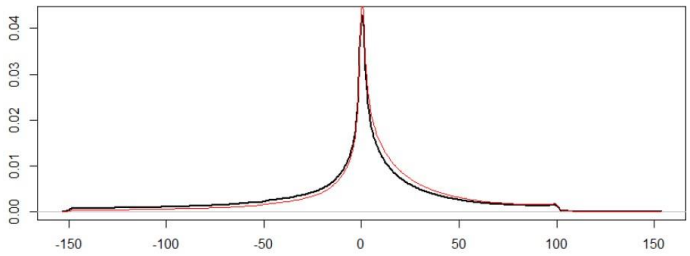

Fig 1.4 Comparazione densità prima (nera) e dopo (rossa) sostituzione missing per la variabile ROE

Quello che è stato possibile notare, performando anche dei Welch t-test è che le variabili con pochi missing values hanno visto un'attribuzione adeguata dei valori mancanti. Viceversa, invece, le variabili con molti missing (percentuali anche sopra il 20%) hanno riscontrato un'efficienza minore nella sostituzione, che comunque non stravolge completamente la densità della *features* come mostrato nella figura 1.4. Per proseguire l'analisi è stato dunque scelto il primo dataset creato dall'algoritmo.

Trattati i valori mancanti, l'attenzione è stata rivolta agli outliers. La libreria utilizzata è

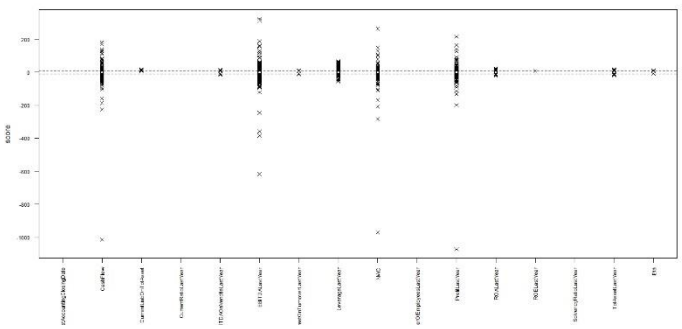

Fig 1.5(a) Outliers microaziende

"**Outforest**" che utilizza *Random Forest* al fine di scovare gli outliers. Essendo il dataset composto da aziende di varie dimensioni, è stato ritenuto opportuno suddividere il dataset in base

alle dimensioni delle stesse e andare ad effettuare un'analisi sugli outliers specifica per ogni grandezza. Nonostante tale accorgimento, il numero di valori che si discostano in maniera significativa dalla distribuzione era elevato. A tal proposito di è deciso di considerare outliers solo quei valori che superavano una treshold di 5 e tra essi si è sostituito il top 1% con la metodologia "**pmm**". Le variabili *Ateco e TaxCodeNumber* sono state escluse dall'analisi. Nelle figure 1.5 sono stati rappresentati gli outliers per ogni variabile implicata nel processo considerando i dati scalati.

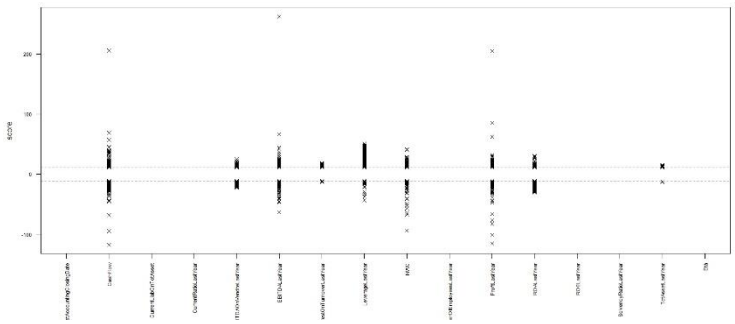

Fig 1.5(b) Outliers piccolo aziende

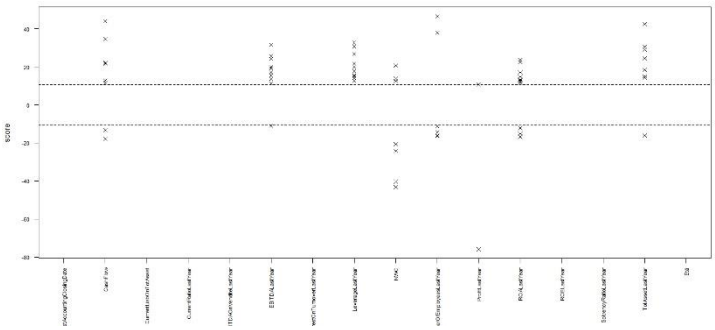

Fig 1.5(c) Outliers aziende medio grandi

1.2 Ulteriore elaborazione dei dati ai fini comparativi

Solo per eseguire le ricerche di tipo comparativo, è stato deciso di raggruppare alcune variabili in maniera tale da cambiarne la granularità. Le alterazioni effettuate sono le seguenti:

- **ATECO** : I vari codici Ateco sono stati raggruppati in sezioni (dalla A alla U) seguendo quelle che sono le disposizioni Istat e poi a loro volta sono state raggruppate andando a formare i seguenti gruppi: A BCDE, F, GHI, J, K, L, MN, OPQ, RSTU.
- Le varie **regioni** sono state raggruppate in 4 categorie secondo tale distinzione:
    - Italia Settentrionale = Emilia-Romagna, Friuli Venezia Giulia, Liguria, Lombardia, Piemonte, Trentino-Alto-Adige, Valle D'Aosta, Veneto.
    - Italia Centrale: Lazio, Marche, Toscana, Umbria.
    - Italia Meridionale: Abruzzo, Basilicata, Campania, Calabria, Molise, Puglia.
    - Italia Insulare: Sardegna, Sicilia.
- La **forma sociale** delle aziende *Association, Foundation, Public Agency, S.C.A.R.I., S.A.S., Mutual Aid Society, S.A.P.A., Foreign Company, S.N.C.*, sono state attribuite ad *Other* in quanto presentavano poche osservazioni e rendevano molto difficile un'analisi statistica soddisfacente.
- Per poter comparare la densità della variabile (DimensioneAzienda), si è considerato il log della variabile Tot Asset in accordo con le fonti ritrovate in letteratura.[1]
- L'età è stata divisa in bin 0-1, 2-4, 5-7, 8-10, 10-15, 16-20, 21-30, 31-40, 41-60, 61-80, 81-122

2. Comparazione fra aziende attive e fallite nell'anno di rendicontazione 2018

In questa sezione, teniamo in considerazione solo di quelle aziende, che hanno redatto bilancio nell'anno 2018. La scelta di tale anno non è stata casuale, ma è dovuta alla presenza di maggior dati per i quale predisporre l'analisi. Il numero di aziende attive è di 795613, mentre quelle fallite è di 75316. Come punto di partenza è stata comparata la densità delle variabili *Età* e *Log_Asset* tenendo in considerazione come fattore discriminante il loro status (Figure 2.1).

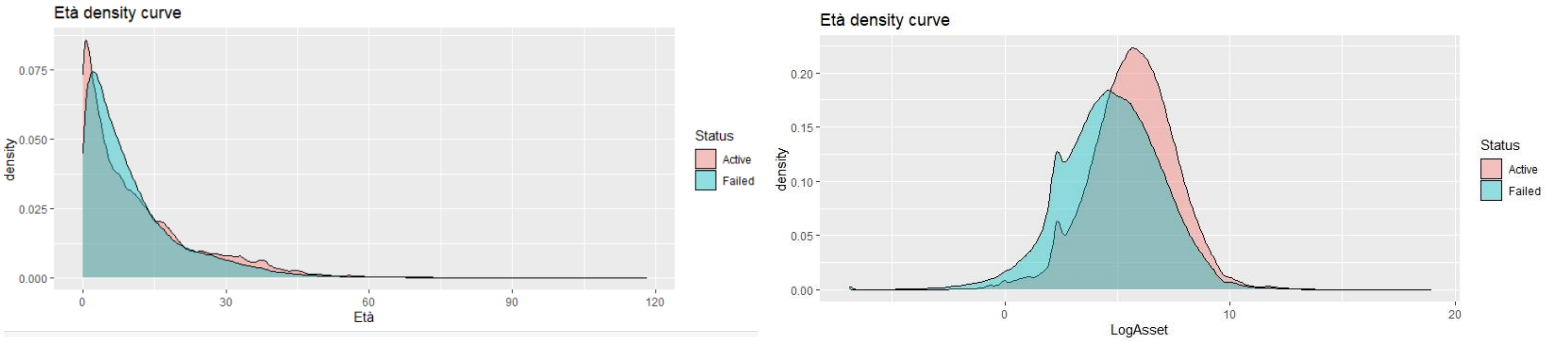

Figura 2.1(a) Densità di aziende attive e fallite a confronto per età e dimensione

Per verificare se le distribuzioni possono essere considerate simili, si è prima utilizzato il ***Anderson–Darling*** *normality test*. Essendo risultate per entrambi le sezioni delle variabili rigetta la normalità, si è utilizzato il ***Welch t-test*** per verificare se le medie delle due distribuzioni si possano ritenere statisticamente simili oppure no. Per entrambi viene rigettata l'ipotesi nulla $\mu_{Active} = \mu_{Failed}$. L'Età media delle aziende fallite risulta essere più alta dell'età media delle aziende attive (14,257 > 13,462). Per quanto riguarda la dimensione invece, le imprese fallite tendono ad avere una dimensione inferiore a quelle attive (4.613 < 5.971).

Spostando l'attenzione invece su forma legale, settore economico e provenienza geografica, si è andati ad analizzare la distribuzione delle aziende attive e fallite nel 2018. Nella figura 2.2 è riportato a scopo esemplificativo il barplot della variabile *Legal form*. Come si può notare, nel 2018 i consorzi e le Scarl presentano il maggior valore percentuale di aziende fallite (rispettivamente

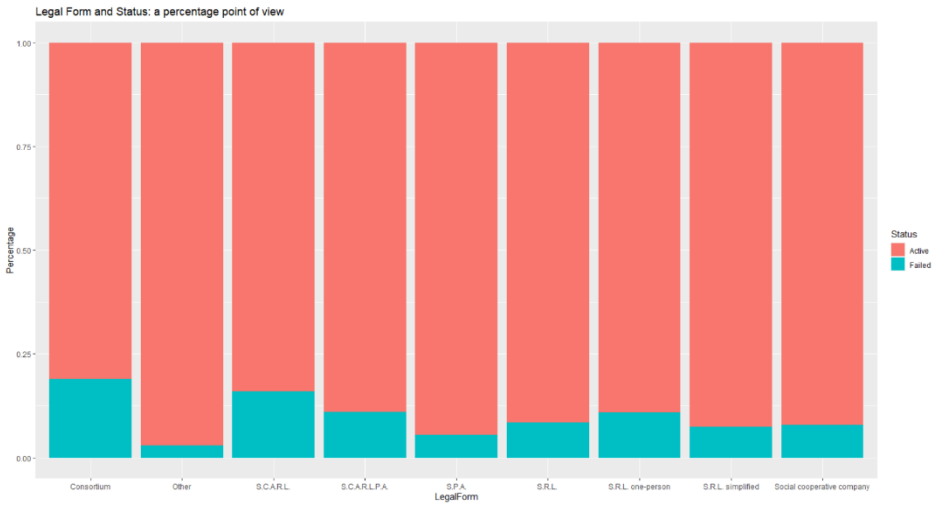

Fig 2.2 Comparazione fra attive e fallite in base alla forma legale.

18,84% e 15,47%), mentre le *Other* sono quelle associate ad una più bassa percentuale di fallimento. Considerando invece area di provenienza e settore economico, non vi sono particolari differenze fra le varie parti d'Italia e i vari settori economici.

2.1 Comparazione delle distribuzioni delle aziende attive e fallite considerando età e dimensione al variare della forma legale e del gruppo Ateco

Dopo aver dato un'impostazione generale all'analisi, verrà svolta una comparazione più specifica che tenga in considerazione non solo dell'età e della dimensione delle aziende, ma anche di come la distribuzione di esse possa variare se si considerano la loro forma legale, la dimensione e la loro provenienza geografica. Il modus operandi di tale ricerca è stato il seguente per ogni coppia di distribuzioni:

- *Anderson-Darling normality test:* H_0 = I dati seguono una distribuzione normale
  H_1 = I dati non seguono una distribuzione normale
- Se AD test accetta H_0 allora si esegue un *F-Test* per verificare se le varianze sono omoschedastiche (H_o) o se non lo sono (H_1)
- Se F-Test accetta H_0 allora si utilizza *t-test* per verificare se $\mu_0 = \mu_1$ (H_0) oppure se $\mu_0 \neq \mu_1$ (H_1)
- Se F-Test non accetta H_0 o una delle due variabili non è distribuita normalmente, allora si esegue un *Welch t-test* per verificare se $\mu_0 = \mu_1$ (H_0) oppure se $\mu_0 \neq \mu_1$ (H_1)

Tutti i test sono considerati accettati per un p-value maggiore o uguale al 5%. Data l'ingente mole di dati, riportiamo esclusivamente un riassunto delle performance. Partiamo considerando la dimensione.

Comparazione distribuzioni aziende attive e fallite considerando la dimensione e le loro caratteristiche sottostanti				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0
S.r.l	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Rigetto H_0
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Rigetto H_0
Other	Rigetto H_0	---	---	Rigetto H_0
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0
GRUPPO ATECO		---	---	
A	Rigetto H_0	---	---	Rigetto H_0
BCDE	Rigetto H_0	---	---	Rigetto H_0
F	Rigetto H_0	---	---	Rigetto H_0
GHI	Rigetto H_0	---	---	Rigetto H_0
J	Rigetto H_0	---	---	Rigetto H_0
K	Rigetto H_0	---	---	Rigetto H_0
L	Rigetto H_0	---	---	Rigetto H_0
MN	Rigetto H_0	---	---	Rigetto H_0
OPQ	Rigetto H_0	---	---	Rigetto H_0
RSTU	Rigetto H_0	---	---	Rigetto H_0

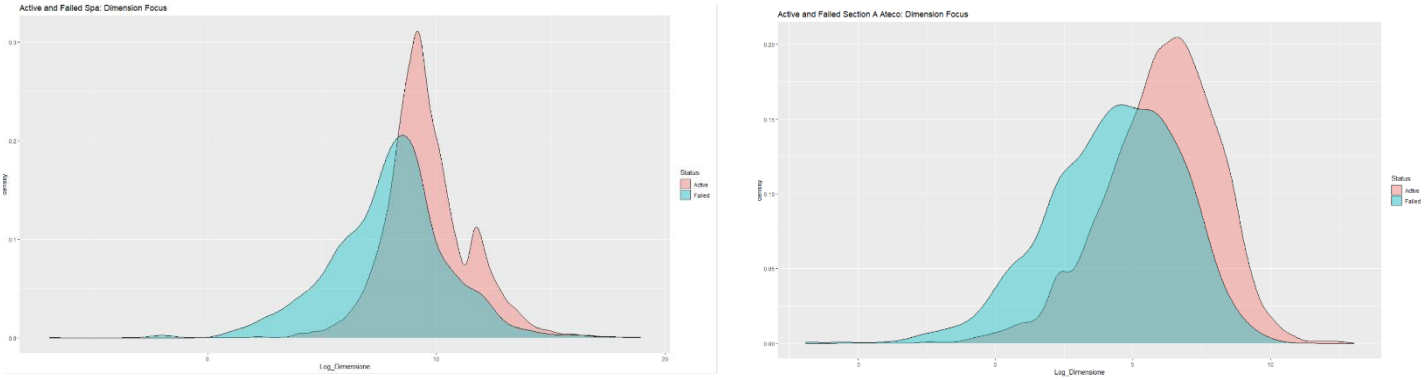

Fig2.3 Densità di S.p.a e Gruppo Ateco A considerando la log_dimensione messe a confronto

Come è possibile evincere dalla tabella riportata pocanzi, non è possibile trovare nessuna prova statisticamente valida che possa confermare il fatto che le distribuzioni di aziende fallite e attive siano similmente distribuite. È possibile quindi dire che le caratteristiche delle aziende in base anche alla loro dimensione influiscono positivamente o negativamente sulla possibilità di fallire.

Andiamo adesso ad analizzare l'età in maniera più approfondita.

Comparazione distribuzioni aziende attive e fallite considerando l'età e le loro caratteristiche sottostanti				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0
S.r.l	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Rigetto H_0 (0.0024)
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Rigetto H_0
Other	Rigetto H_0	---	---	Accetto H_0 (0.7733)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0
GRUPPO ATECO		---	---	
A	Rigetto H_0	---	---	Rigetto H_0
BCDE	Rigetto H_0	---	---	Rigetto H_0
F	Rigetto H_0	---	---	Rigetto H_0
GHI	Rigetto H_0	---	---	Rigetto H_0 (0.0011)
J	Rigetto H_0	---	---	Rigetto H_0 (0.0457)
K	Rigetto H_0	---	---	Rigetto H_0
L	Rigetto H_0	---	---	Rigetto H_0
MN	Rigetto H_0	---	---	Rigetto H_0
OPQ	Rigetto H_0	---	---	Rigetto H_0
RSTU	Rigetto H_0	---	---	Accetto H_0 (0.167)

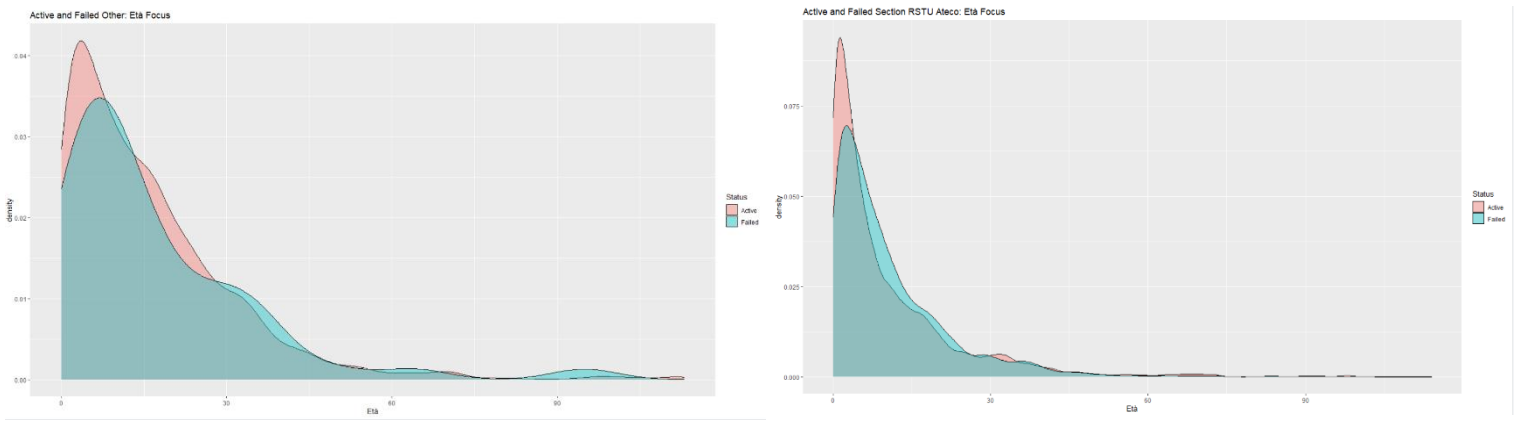

Fig2.4 Densità di Other e Gruppo Ateco RSTU considerando l'età messe a confronto

A differenza dell'analisi sulla dimensione, dove non si sono apprezzate differenze, per l'età invece è statisticamente possibile provare che, prendendo in considerazione la categoria *Other* per il subset *Forma legale* e il gruppo *RSTU* (che varia dalle attività artistiche a quelle di altri servizi, passando da quelle di finalità domestica a quelle di organismi extraterritoriali) per il *subset Ateco* le variazioni siano dovute al caso. In questa casistica, infatti, appartenere ad uno di questi sottogruppi non permettere in media di delineare profili differenti fra aziende fallite e non. Nonostante sia stata rigettata l'ipotesi nulla, è comunque interessante sottolineare come il gruppo *Ateco GHI* riporti valori di p-value che è possibile accettare con test ad esempio al 99%..

3. Comparazione fra aziende fallite negli anni di rendicontazione 2016-2018

La stessa analisi che è stata svolta per l'anno 2018 viene invece adesso svolta considerando se è possibile trovare delle similarità fra le distribuzioni di aziende fallite negli anni 2016-2018. Come prima prenderemo prima in esamine solo la distribuzione dell'età e della dimensione, per poi invece andarla a particolareggiare con le varie caratteristiche dell'imprese. Per completezza ripetiamo quanto riportato già prima, in quanto l'iter procedurale è il medesimo:

- *Anderson-Darling normality test:* H_0 = I dati seguono una distribuzione normale
  H_1 = I dati non seguono una distribuzione normale
- Se AD test accetta H_0 allora si esegue un *F-Test* per verificare se le varianze sono omoschedastiche (H_o) o se non lo sono (H_1)
- Se F-Test accetta H_0 allora si utilizza *t-test* per verificare se $\mu_0 = \mu_1$ (H_0) oppure se $\mu_0 \neq \mu_1$ (H_1)
- Se F-Test non accetta H_0 o una delle due variabili non è distribuita normalmente, allora si esegue un *Welch t-test* per verificare se $\mu_0 = \mu_1$ (H_0) oppure se $\mu_0 \neq \mu_1$ (H_1)

Anche in tale caso si accetta l'ipotesi nulla solo per valori di p-value maggiori o uguali al 5%.

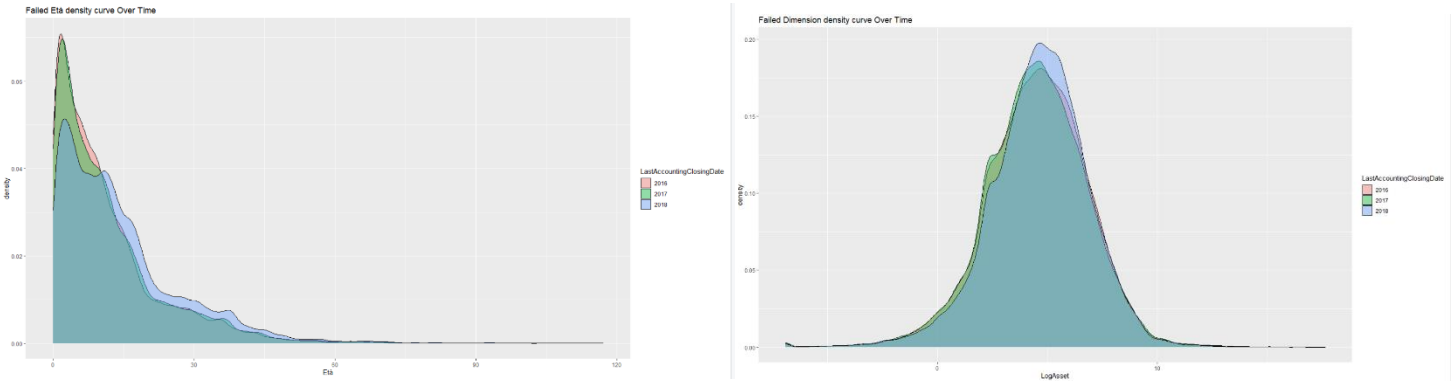

Fig 3.1 Densità di Età e Log_Dimensione prendendo in considerazione gli anni 2016_2018 e le aziende fallite

Tutti i test rifiutano l'ipotesi nulla. Questo, come visto in precedenza, significa che i fallimenti nel corso degli anni delle aziende comparati per età o dimensione non sono similarmente distribuiti e dunque non sono attribuibili al caso. Una comparazione è stata fatta anche per le variabili *Legal form, GroupedRegions, Ateco,* non mostrando comunque essenziali differenze.

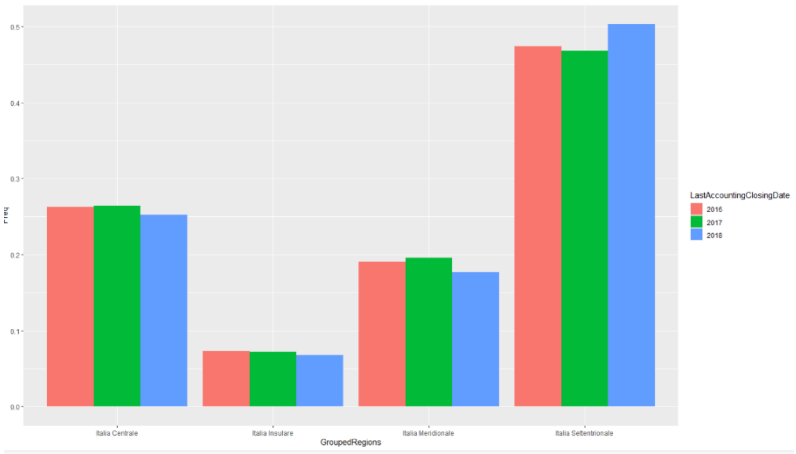

Fig 3.2 Comparazione delle aziende fallite in base alla provenienza geografica

3.1 Comparazione delle distribuzioni delle aziende attive e fallite considerando età e dimensione al variare della forma legale e della provenienza geografica.

Vengono adesso riportati i risultati delle analisi per la comparazione più approfondita della distribuzione delle aziende fallite nei 3 anni. A causa della mole di dati elaborati, riportiamo esclusivamente le tavelle che riassumono le performance dei vari test riportando il p-value quando necessario.

Iniziamo con la comparazione tra le aziende nel 2016 e nel 2017 tenendo in considerazione la loro dimensione e la loro provenienza geografica.

Comparazione distribuzioni aziende fallite considerando la dimensione e le loro caratteristiche sottostanti tra 2016 e 2017				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Accetto H_0 (0.4939)
S.r.l	Rigetto H_0	---	---	Accetto H_0 (0.3988)
S.r.l one person	Rigetto H_0	---	---	Accetto H_0 (0.2206)
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Accetto H_0 (0.3932)
Other	Rigetto H_0	---	---	Accetto H_0 (0.728)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Accetto H_0 (0.1592)
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0 (0.015)
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Rigetto H_0
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0 (0.0496)
Italia Meridionale	Rigetto H_0	---	---	Accetto H_0 (0.789)
Italia Insulare	Rigetto H_0	---	---	Accetto H_0 (0.2484)

Proponiamo adesso la stessa analisi per gli anni 2016 e 2018.

Comparazione distribuzioni aziende fallite considerando la dimensione e le loro caratteristiche sottostanti tra 2016 e 2018				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0
S.r.l	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Accetto H_0 (0.643)
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Accetto H_0	Rigetto H_0	---	Accetto H_0 (0.09)
Other	Rigetto H_0	---	---	Accetto H_0 (0.2975)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Accetto H_0 (0.088)
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Accetto H_0 (0.1143)
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0 (0.04)
Italia Meridionale	Rigetto H_0	---	---	Rigetto H_0
Italia Insulare	Rigetto H_0	---	---	Rigetto H_0

Proponiamo adesso la stessa analisi per gli anni 2017 e 2018.

Comparazione distribuzioni aziende fallite considerando la dimensione e le loro caratteristiche sottostanti tra 2017 e 2018				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0
S.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Accetto H_0 (0.077)
s.r.l simplified	Rigetto H_0	---	---	Accetto H_0 (0.3912)
Social cooperative	Rigetto H_0	---	---	Accetto H_0 (0.4697)
Other	Accetto H_0	Accetto H_0	Accetto H_0 (0.207)	---
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Rigetto H_0
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0
Italia Meridionale	Rigetto H_0	---	---	Rigetto H_0
Italia Insulare	Rigetto H_0	---	---	Rigetto H_0

Come è possibile notare dalle tabelle, rispetto all'analisi generale dove le ipotesi nulle sono state rigettate con forza, in questo caso se ciò non avviene. È possibile dunque affermare che, soprattutto nel sottogruppo legal form e fra gli anni 2016, 2018 e 2016, 2017 ci sia una certa somiglianza fra la distribuzione delle aziende fallite considerandone le caratteristiche in esame. Questo ci suggerisce che se dovessimo analizzare le aziende fallite in tali anni, sotto gli aspetti sopra esaminati non

troveremo in media grandi differenze, andando quindi a rendere difficile una possibile classificazione.

Iniziamo con la comparazione tra le aziende nel 2016 e nel 2017 tenendo in considerazione la loro età e la loro forma sociale.

Comparazione distribuzioni aziende fallite considerando la dimensione e le loro caratteristiche sottostanti tra 2016 e 2017				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Accetto H_0 (0.5588)
S.r.l	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Rigetto H_0
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Accetto H_0 (0.7895)
Other	Rigetto H_0	---	---	Accetto H_0 (0.5718)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Accetto H_0(0.085)
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Accetto H_0(0.5851)
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0 (0.0421)
Italia Meridionale	Rigetto H_0	---	---	Rigetto H_0
Italia Insulare	Rigetto H_0	---	---	Rigetto H_0

Proponiamo adesso la stessa analisi per gli anni 2016 e 2018.

Comparazione distribuzioni aziende fallite considerando l'età e le loro caratteristiche sottostanti tra 2016 e 2018				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0
S.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Rigetto H_0
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Rigetto H_0
Other	Rigetto H_0	---	---	Accetto H_0 (0.3447)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Rigetto H_0
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0
Italia Meridionale	Rigetto H_0	---	---	Rigetto H_0
Italia Insulare	Rigetto H_0	---	---	Rigetto H_0

Proponiamo adesso la stessa analisi per gli anni 2017 e 2018.

Comparazione distribuzioni aziende fallite considerando l'età e le loro caratteristiche sottostanti tra 2017 e 2018				
Variabile	**Anderson-Darling**	**F-Test**	**T-Test**	**Welch T-Test**
LEGAL FORM		---	---	
S.p.a	Rigetto H_0	---	---	Rigetto H_0(0.011)
S.r.l.	Rigetto H_0	---	---	Rigetto H_0
S.r.l one person	Rigetto H_0	---	---	Rigetto H_0
s.r.l simplified	Rigetto H_0	---	---	Rigetto H_0
Social cooperative	Rigetto H_0	---	---	Accetto H_0 (0.5352
Other	Rigetto H_0	---	---	Accetto H_0 (0.23)
Consortium	Rigetto H_0	---	---	Rigetto H_0
S.c.a.r.l.	Rigetto H_0	---	---	Accetto H_0 (0.0507)
S.c.a.r.l.p.a.	Rigetto H_0	---	---	Rigetto H_0
GROUPED REGIONS		---	---	
Italia Settentrionale	Rigetto H_0	---	---	Rigetto H_0
Italia Centrale	Rigetto H_0	---	---	Rigetto H_0
Italia Meridionale	Rigetto H_0	---	---	Rigetto H_0
Italia Insulare	Rigetto H_0	---	---	Rigetto H_0

A differenza di quanto visto quando si è analizzato la dimensione rispetto alle varie forme legali e provenienza geografiche, per l'età delle aziende la maggior parte delle distribuzioni suddivise poi nelle varie forme e regioni non mostrano similarità. Soprattutto nella comparazione fra 2016 e 2018, solo le aziende nella categoria *Other* hanno una distribuzione dei fallimenti simili. In aggiunta, *Other* è l'unica categoria che in tutti gli anni di comparazione mostra una similarità fra le distribuzioni dei fallimenti. Le aziende sociali cooperative mostrano invece grandi similarità fra gli anni 2016, 2017 e 2017, 2018.

4. Analisi della distribuzione dei fallimenti nel 2018

Dopo aver approfondito e analizzato le distribuzioni delle variabili di interesse attraverso il tempo e attraverso alcune caratteristiche specifiche delle imprese, ci chiediamo come i fallimenti si siano distribuiti nell'arco dell'anno 2018 considerando alcune variabili.

Considerando l'età, la probabilità di fallire più alta si ha per le aziende che sono sul mercato da non molti anni. L'apice è raggiunto per le aziende che sono sul mercato dagli 8 ai 20 anni, con percentuali che si aggirano intorno al 10%. La probabilità più bassa di fallire si riscontra comunque nel primo anno di vita dell'azienda (5,33%).

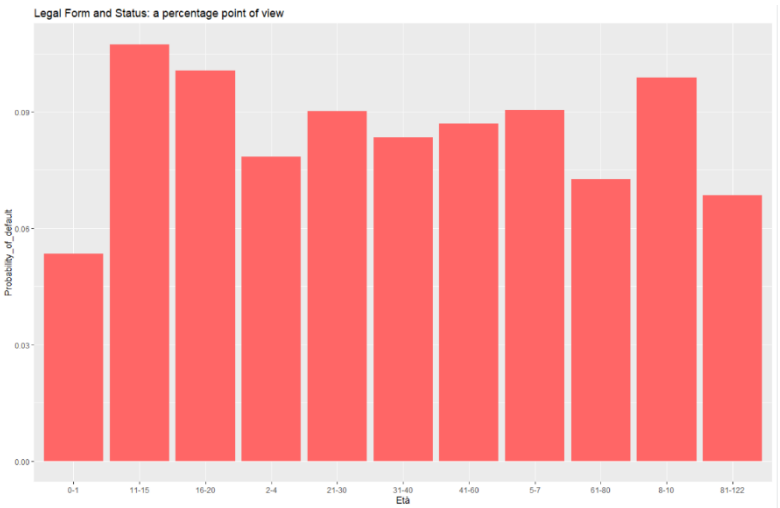

Fig 4.1 Probabilità di fallire nel 2018 condizionata all'età dell'impresa

Per quanto riguarda invece la probabilità di fallimento condizionata alla dimensione, le microaziende sono quelle che presentano la maggior probabilità di fallimento (9.6%), con le piccole medio imprese che invece si attestano su valori che vanno dal 2.6% al 3%.

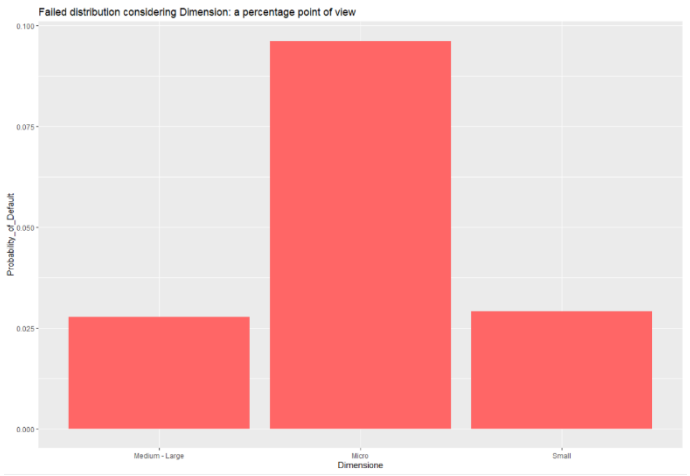

Fig 4.2 Probabilità di fallire nel 2018 condizionata alla dimensione dell'impresa

Al fine di svolgere un'analisi maggiormente dettagliata, è stato deciso di esaminare non solo la probabilità condizionata per età e dimensione, ma come fatto in precedenza, sono state prese in considerazione eventuali cambiamenti riguardanti le caratteristiche delle aziende. Dall'analisi svolta è risultato quanto segue.

Partendo dalla provenienza geografica delle aziende, è stato visto che per la regione settentrionale sia per l'età, sia per la dimensione, le probabilità di fallimento rispecchiano quelle nazionali. Nella parte centrale dell'Italia invece si è rilevato una maggiore probabilità di fallire per quanto riguarda le aziende in quasi tutte le sue fasce di età, mentre non si hanno particolari distaccamenti per le dimensioni. Anche nell'Italia meridionale si ha una situazione simile, con il picco della probabilità di fallimento raggiunta nell'età 61-80 (12,54%). Regolare invece la situazione per la dimensione. Per quanto riguarda invece le isole, anche qua si registra un picco di fallimento in età più avanzata (12, 44% tra i 41-60 anni), mentre si registra anche un'incrementata probabilità di fallimento per le imprese medio grandi passando dal 2,76% al 4,49%.

Passando al Gruppo Ateco, tra le aziende appartenenti al gruppo A (agricoltura silvicoltura e pesca) vi è un generalizzato ribasso di probabilità di fallimento data l'età. Soprattutto fra i 16 e i 40 anni essa si aggira sul 7%, registrando un ribasso dai 2 ai 3 punti percentuali. Situazione simile si ha per le aziende del gruppo L (attività immobiliari) e per le aziende di estrazione mineraria, manifatturiere, di fornitura di energia ed acqua (BCDE). Situazione diversa si ha invece per le aziende che trattano l'ambito delle costruzioni. In questo caso si registra un'aumentata probabilità di fallire dopo i 10 anni con un picco del 13,41% tra i 41 e i 60 anni. Anche la probabilità di fallimento condizionata la dimensione presenta valori più alti per le microaziende, mentre registra un calo per quelle di medio grande dimensione. Situazione simile viene vissuta dal gruppo MN che si occupa di attività scientifiche, professionali e di noleggio o viaggi. In questo caso, oltre ad un aumento della probabilità di fallimento soprattutto tra gli 8 e i 10 anni, si registra anche un aumento per le imprese di piccola dimensione. Presentano invece valori molto più bassi le aziende che si occupano di sanità, istruzione e sicurezza. La probabilità di fallire in ogni fascia della propria vita aziendale è molto inferiore rispetto alla media nazionale, riproponendo la stessa tendenza anche quando si tratta di dimensione dell'impresa. Le restanti categorie (RSTU, L, GHI) presentano più o meno caratteristiche riferibile alla media nazionale, eccezione fatta per RSTU che presenta un picco di fallimento nei primi anni di vita (12% tra gli 8 e i 10 anni).

Trattando invece la forma giuridica dell'azienda, possiamo subito affermare che per le S.r.l. in tutte le loro forme, la probabilità dei fallimenti segue quella nazionale. Questo non sorprende molto in quanto esse rappresentano gran parte del dataset. È stata riscontrata invece un marcato ribasso della probabilità di fallire nelle società per azioni (Fig 4.3), soprattutto nella fascia di età 31-40. Sulla stessa scia troviamo anche le aziende che sono etichettate nella classe *other*. Vi sono invece forme giuridiche come i consorzi e le S.c.a.r.l. che presentano marcate differenze in negativo rispetto alla popolazione generale. Nei consorzi, infatti, la probabilità di fallimento per le di età dagli 11 ai 40 anni supera il 20% con un picco tra i 31 e i 40 del 24,77%. Stessa situazione si ha facendo un focus sulla dimensione, dove i micro-consorzi hanno una probabilità di fallire del 2018 doppia rispetto alla media nazionale delle microaziende. Come si diceva anche le S.c.a.r.l. presentano caratteristiche simili, ma più improntate sui primi anni di vita dell'azienda. Si ha infatti una probabilità di fallimento che cresce nei primi anni di vita fino ad arrivare al picco tra gli 8 e i 10 anni del 26,58%, per poi cominciare una lenta discesa verso valori, comunque, sopra la media nazionale. Infine, le S.c.a.r.l.p.a. presentano valori leggermente più alti della media nazionale per quanto riguarda i fallimenti soprattutto quando si trattano le fasce di età. viceversa le aziende sociali cooperative mostrano alcuni dei valori di fallimento più bassi fra i 31 e i 60 anni con valori intorno al 3-4%.

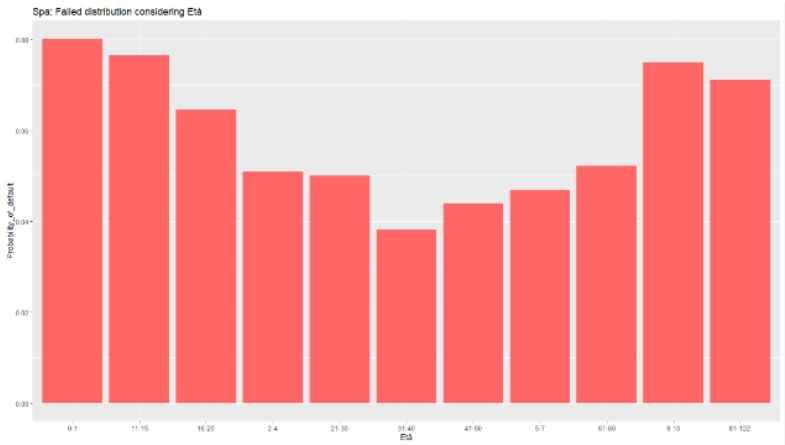

Fig 4.3 Probabilità di fallire nel 2018 rispetto all'età delle S.p.a.

5. Modelli per la predizione dei fallimenti

Nell'ultima parte dell'analisi è di interesse la costruzione di modelli parametrici e non che possano predire quando un'azienda possa fallire. Per allenare i modelli è stato preso come train il dataset contenente tutte le variabili dell'azienda che si riferivano al penultimo anno di rendicontazione, mentre è stato usato come test set quello con l'ultimo anno di bilancio.

Il primo modello scelto che viene implementato è la **regressione logistica.**

Per prima cosa anche il train è stato trattato come per il test in precedenza per sostituire adeguatamente i missing values e gli outliers. Per evitare di essere ripetitivo ometterò la descrizione del processo che è già stato spiegato in precedenza. Anche in questo caso, infatti, sono state utilizzate le librerie mice e outForest.

Una volta completato questo step è stata resa necessaria un'operazione di selezione delle variabili. Essa si è basata sia su argomentazione trovate in letteratura, sia sulla necessità di evitare al massimo la multicollinearità. A tal proposito è stato indagata la correlazione fra le variabili all'interno del dataset. Per farlo è stato deciso di utilizzare la metodologia di ***Spearman***. Infatti dopo una valutazione riguardo la possibile normalità delle distribuzioni delle variabili (tramite Anderson-darling test) è stato semplice concludere che le variabili non si comportavano come una normale bivariata. Le variabili che sono state quindi eliminate presentavano con le altre un grado di

correlazione superiore a 0.7. Le variabili rimosse dal train sono dunque: *Roa, Net working capital ed ebitda, profit.* In aggiunta a tali variabili, sono state eliminate le variabili: *Last accounting closing date, Legal form, Registered office address, Ateco, Dimensione azienda e Tax code number.* Successivamente si è ricalcolata la correlazione tra le variabili che ha portato i risultati in figura 5.1.

Una volta selezionate le features è stato dunque implementato il modello. Il modello mostra che nessuna delle variabili è statisticamente differente da 0, non rendendo quindi necessaria su tale base una successiva rielaborazione del gruppo delle variabili. Avendo comunque timore che vi sia ancora presenza di multicollinearità o di variabili superflue

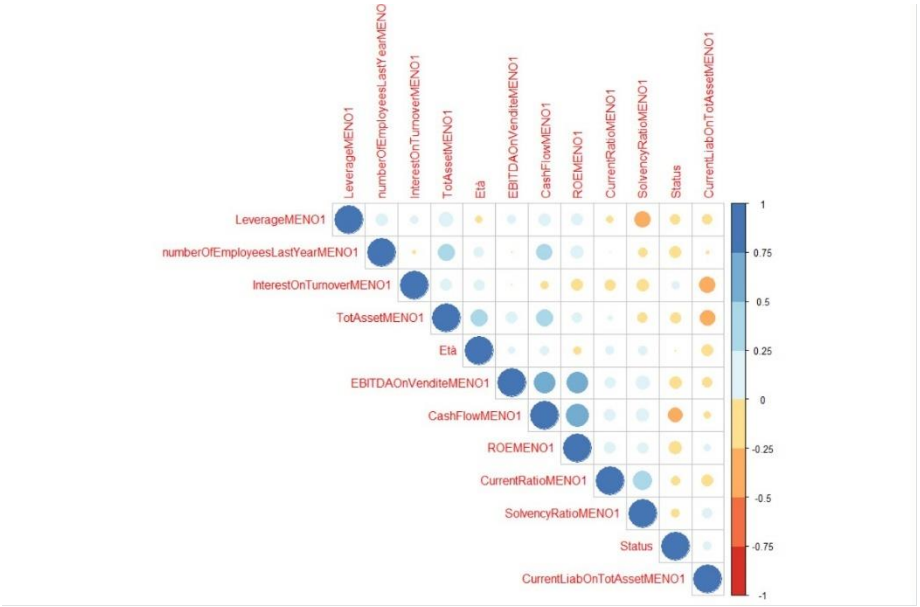

Fig 5.1 Matrice di correlazione tra le variabili del train dopo la selezione

all'interno del gruppo delle predittori, è stato indagato anche il *Variance inflaction factor* ed è stato implementato l'algoritmo *StepAIC.* Entrambi però non hanno segnalato possibili migliorie all'interno di tale gruppo. Per i vari coefficienti sono stati calcolati anche i relativi intervalli di confidenza al 95%

	2.5 %	97.5 %
(Intercept)	-1.059285e+00	-1.034671e+00
CashFlow	-6.002054e-06	-2.916483e-06
CurrentLiabOnTotAsset	6.238967e-01	6.480668e-01
CurrentRatio	1.837004e-02	2.271838e-02
EBITDAOnVendite	-1.602787e-03	-1.525283e-03
InterestOnTurnover	1.203725e-02	1.276988e-02
Leverage	-1.456932e-04	-6.636172e-05
numberOfEmployees	-1.806585e-03	-1.454679e-03
ROE	-8.017984e-03	-7.852084e-03
SolvencyRatio	-3.307701e-03	-3.070958e-03
TotAsset	3.192734e-08	8.616399e-08
Età	-7.515691e-03	-6.965800e-03

```
glm(formula = Status ~ ., family = binomial("logit"), data = Train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1599  -0.9189  -0.7936   1.3238   6.4201

Coefficients:
                       Estimate Std. Error  z value Pr(>|z|)
(Intercept)          -1.047e+00  6.279e-03 -166.739  < 2e-16 ***
CashFlow             -4.410e-06  7.882e-07   -5.596 2.20e-08 ***
CurrentLiabOnTotAsset  6.360e-01  6.166e-03  103.143  < 2e-16 ***
CurrentRatio          2.054e-02  1.109e-03   18.521  < 2e-16 ***
EBITDAOnVendite      -1.564e-03  1.977e-05  -79.102  < 2e-16 ***
InterestOnTurnover    1.240e-02  1.869e-04   66.365  < 2e-16 ***
Leverage             -1.059e-04  2.024e-05   -5.234 1.66e-07 ***
numberOfEmployees    -1.628e-03  8.978e-05  -18.132  < 2e-16 ***
ROE                  -7.935e-03  4.232e-05 -187.491  < 2e-16 ***
SolvencyRatio        -3.189e-03  6.039e-05  -52.808  < 2e-16 ***
TotAsset              5.784e-08  1.397e-08    4.142 3.45e-05 ***
Età                  -7.241e-03  1.403e-04  -51.615  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2249208  on 1728706  degrees of freedom
Residual deviance: 2162588  on 1728695  degrees of freedom
AIC: 2162612
```

L'unico valore tra quelli calcolati che permette di avere un odds sostanzialmente diverso da 1 (1,88) è *CurrentLiabilitiesOnTotalAsseth* che risulta essere dunque una delle variabili più discriminatorie del modello,

A questo punto è stata calcolata la **performance** del nostro classificatore prima sul train e poi sul test set. Il livello di accuratezza raggiunto sul train è del 65,6%, mentre sul test set è del 67,15%. Le performance non sono dunque molto soddisfacenti, seppur risultano essere leggermente preferite al *No information rate,* andando anche a rigettare l'ipotesi nulla che esso sia statisticamente migliore dell'accuratezza della regressione logistica.

Successivamente, oltre ai risultati dell'accuratezza, è stata implementata anche la **Roc curve** e il **Calibration plot**.

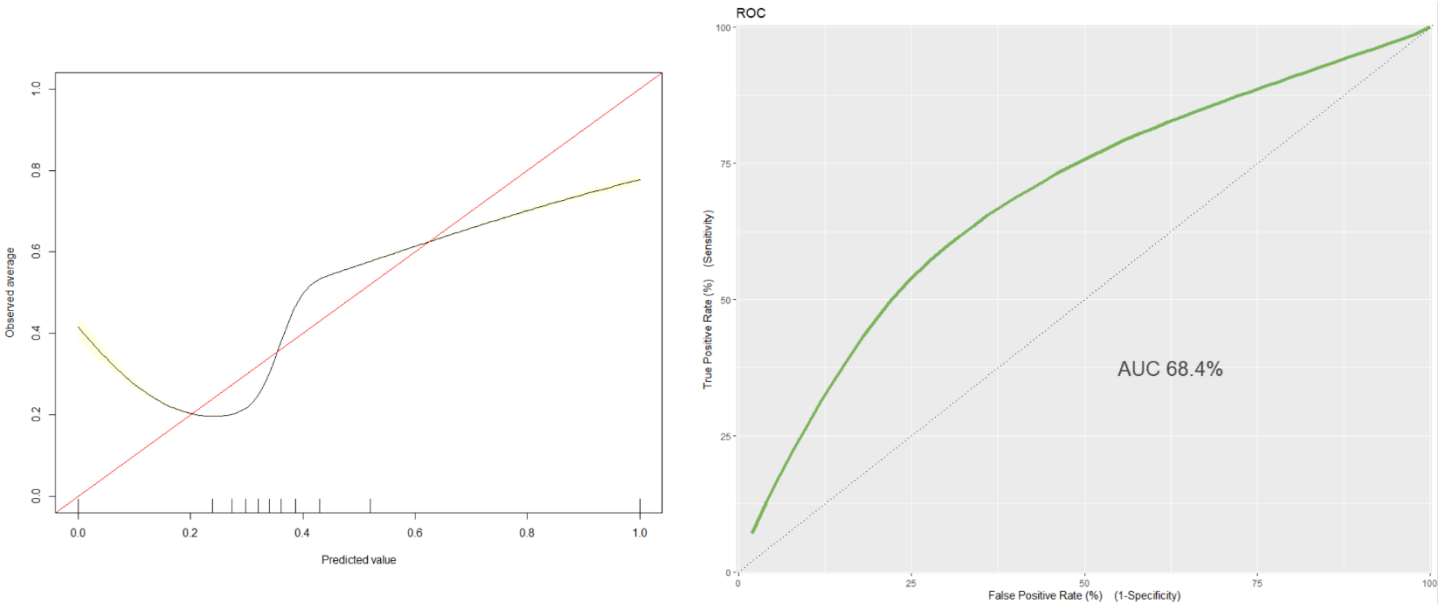

Fig 5.2 Calibration plot e Roc curve della regressione logistica

Il calibration plot e la roc curve ci confermano i risultati poco promettenti del modello. Nel calibration plot è possibile vedere come in fasi del tutto alternati la probabilità empirica è maggiore di quella da noi predetta (quando la curva supera la diagonale) o minore di quella predetta (quando è sottostante la diagonale). Per avere un modello calibrato ci si aspetterebbe di vedere che la curva generata dal nostro modello segua in maniera più o meno fedele quella ideale della diagonale. Passando invece alla Roc curve, anch'essa conferma le attese. Analizzando la curva abbiamo che raggiungiamo alti livelli di sensibilità solo quando ormai il modello è poco specifico, andando quindi a collezionare molti falsi positivi. Anche l'auc conferma che, seppur il classificatore è migliore di un modello randomico, copre solo il 68,4% dell'area. Infine, è stato utilizzato anche il ***Hosmer-Lemeshow*** *test.* Tale test che si basa sulla Chi-quadro riporta un p-value che se al di sotto della soglia di accettazione indica la non bontà del fitting. Nel caso in questione il p-value porta a rigettare l'ipotesi nulla confermando ancora una volta la performance non ottimale della regressione.

Per questo motivo si è considerata e sviluppata l'idea di provare a migliorare il nostro modello utilizzando la regressione logistica con la penalizzazione ***Lasso.*** Essendo molto importante la definizione del parametro lambda, è stato deciso di utilizzare il modello con CV al fine di ottenere un valore ideale di tale parametro. Il valore scelto per il lambda è dunque 0.0006847788. Una volta fittato il modello, i coefficienti che sono stati trovati sono molto simili a quelli precedenti, ad eccezione che di *TotAsset* che è stata eliminata dal modello. Infatti, come è ben risaputo, il Lasso opera pure con un modello di features selection.

	s0
(Intercept)	-1.028320e+00
CashFlow	-9.741877e-07
CurrentLiabOnTotAsset	6.115952e-01
CurrentRatio	1.575620e-02
EBITDAOnVendite	-1.546511e-03
InterestOnTurnover	1.199893e-02
Leverage	-6.401857e-05
numberOfEmployees	-9.577265e-04
ROE	-7.886945e-03
SolvencyRatio	-2.954326e-03
TotAsset	.
Età	-7.104376e-03

Anche per il lasso si va a valutare l'accuratezza sul train e sul test set. L'accuratezza sul train risulta essere 65,58% mentre sul test è del 67,11%. Anche in questo caso dunque le performance sono piuttosto scadenti.

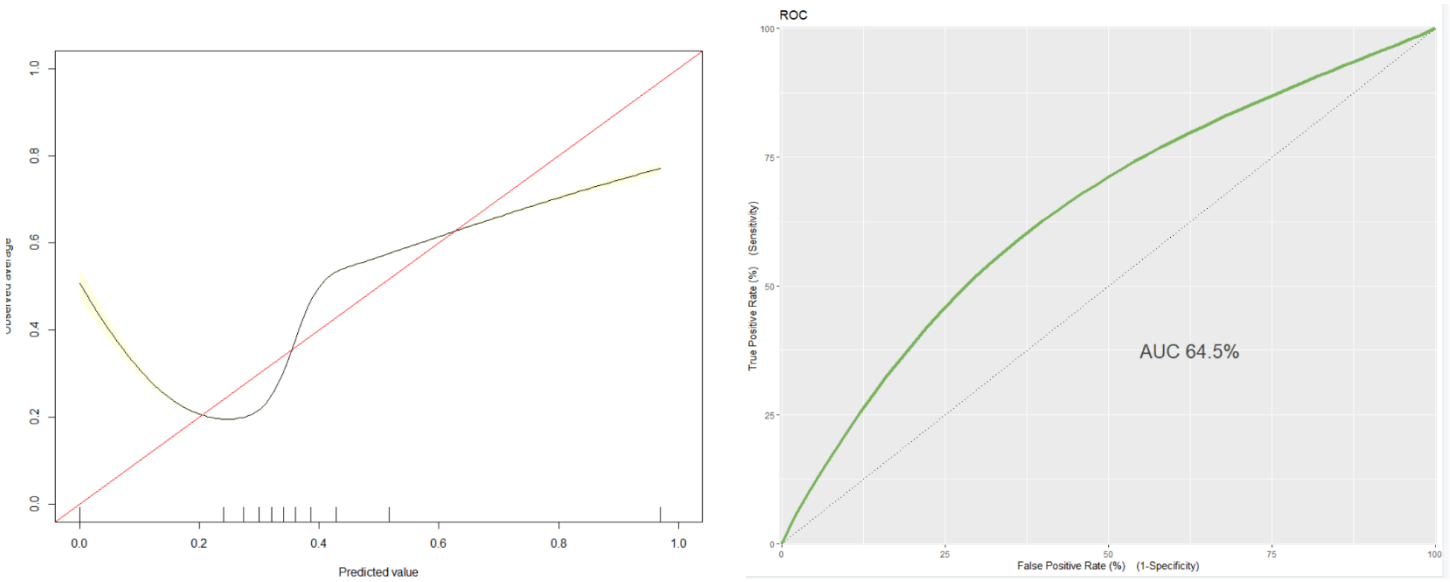

Fig 5.3 Calibration plot e Roc curve della regressione logistica con penalizzazione Lasso

Esaminando il calibration plot e la roc curve del nostro modello con penalizzazione è possibile notare come i due siano molto simili tra loro. In questo caso si evidenzia, seppur di poco, un peggioramento delle condizioni del modello rispetto alla regressione logistica senza penalizzazione.

Alla luce di quanto analizzato fino ad ora, per tale studio, è preferibile la regressione senza penalizzazione rispetto a quella con.

6. Conclusioni

In conclusione, possiamo affermare che, seppur in linea generale la distribuzione tra aziende fallite ed attive risulta essere statisticamente differente, come quella dei fallimenti attraverso i vari anni di rendicontazione, è possibile invece apprezzare alcune similarità nel momento in cui aumentiamo la specificità dell'analisi. Questo consiglia dunque ulteriori analisi riguardo alla determinazione delle variabili che possano essere effettivamente discriminanti per riconoscere in anticipo quei fattori di rischio che, collegati ad un determinato mercato o ad una determinata caratteristiche aziendale, possano portare l'azienda più facilmente a fallire.

Per quanto riguarda i modelli predittivi implementati, la regressione logistica e la sua versione penalizzata con il Lasso non risultano essere sufficienti al fine di determinare in maniera corretta il fallimento o meno di un'azienda.

7. Bibliografia

[1] The influence of variable selection methods on the accuracy of bankruptcy prediction models du Jardin, Philippe Edhec Business School
[2] https://datascienceplus.com/imputing-missing-data-with-r-mice-package/