
# RISK OF BUSINESS FAILURE

**AUTHORS**

Samuele Cucchi, Francesco Falleni, Francesco Gemignani

**Statistical Methods for Data Science**

Data Science & Business Informatics

Academic Year 2020/2021

# Contents

# Introduzione

Il presente elaborato si pone come obiettivo quello di individuare le principali cause che portano al fallimento un'azienda.
Le analisi che abbiamo approfondito riguardano le domande che ci sono state poste: in primo luogo vengono analizzate la distribuzione delle aziende attive e fallite in un determinato anno rispetto alle dimensioni dell'azienda ed alla sua età, valutando l'eventuale eterogeneità in funzione del settore in cui opera (Sito ISTAT) e della forma societaria (i.e. S.R.L., S.P.A.,etc.). In generale, le dimensioni di un'azienda dipendono da differenti parametri: il risultato di esercizio, il numero medio di dipendenti e il totale delle attività. Tutte queste informazioni sono contenute nel bilancio, documento che ogni azienda è tenuta a redigere al termine di ogni esercizio.
Successivamente abbiamo approfondito l'analisi, considerando soltanto le aziende fallite e analizzando l'andamento dell'età e della dimensione in un periodo temporale di 2 o più anni. Anche in questo caso verificando se la distribuzione cambia in funzione della posizione dell'azienda (i.e. regione, provincia, comune,etc.) e della forma societaria.
La terza domanda riguarda, invece, di calcolare la distribuzione di probabilità delle aziende fallite in base all'età e alla dimensione, distinguendo l'andamento in base al settore in cui opera, alla posizione e alla forma societaria.
Infine, abbiamo allenato diversi modelli (regressione logistica (senza regolarizzazione e regolarizzata con Elasticnet) e Random Forest) con l'obiettivo di costruire un classificatore in grado di predire con buone performance se un'azienda è fallita o no.

# 1 Data Understanding & Preparation

Il dataset AIDA che ci è stato fornito contiene le informazioni di $1,894,412$ aziende italiane, ognuna delle quali con 80 features. Alcuni attributi sono presenti per 3 anni consecutivi: le informazioni si possono riferire agli ultimi dati disponibili (last.available.yr), all'anno precedente (-1) o due anni precedenti (-2).

## 1.1 Dataset Assoluto

Abbiamo convertito il dataset originale in un dataset assoluto, nel quale ogni record riporta gli indicatori di una azienda in uno specifico anno, filtrato nella variabile Year. In questo modo la stessa azienda viene identificata in 3 anni. In particolare per ogni record del precedente dataset, vengono creati tre record:

- il primo con i campi relativi all'anno corrente (last.avail)
- il secondo con i campi relativi all'anno precedente (-1)
- il terzo con i campi relativi a due anni precedenti (-2)

Il dataset assoluto è composto da 5,683,236 record e 35 features. Abbiamo utilizzato questo dataset per lo svolgimento dell'intero progetto perchè ci permette di filtrare in modo assoluto l'anno che vogliamo analizzare e riduce la dimensione degli attributi (quelli replicati negli anni).

| | | |
|---|---|---|
| Age ** | EBITDA/Vendite * | Profit.(loss)th * |
| ATECO.2007code | EBITDAth * | Province |
| ATECO.Sector.Name ** | Failed ** | Region |
| Banks/turnover * | Incorporation.year | ROE * |
| Cash.Flowth * | Interest/Turnover * | ROA * |
| Company.name | Legal.form | ROI * |
| Comune.ISTAT.code | Legal.status | ROS * |
| Cost.of.debit * | Leverage * | Solvency.ratio * |
| Current.liabilities/Tot.ass * | Liquidity.ratio * | Tax.code.number |
| Current.ratio * | Net.financial.positionth * | Total.assets.turnover.(times) * |
| Debt/EBITDA.ratio * | Net.working.capitalth * | Total.assetsth * |
| Debt/equity.ratio * | Number.of.employees * | Year ** |

Table 1.1: Tutte le features del dataset AIDA replicate sul dataset assoluto.
(*) variabili assolutizzate ripetute nei 3 precedenti anni (last.avail., -1 e -2),
(**) features che sono state aggiunte al dataset in quanto necessarie per le analisi richieste.

## 1.2 Definizione di Fallimento

Lo stato di un'azienda può essere: Active, Active (default of payments), Active (receivership), Bankruptcy, In Liquidation, Dissolved. Abbiamo definito il concetto di fallimento, prendendo in considerazione tutte le aziende che non sono attive. In modo analogo, avremmo potuto anche scegliere di considerare fallita un azienda che si trovava in stato di bancarotta, ma la prima definizione, oltre ad essere più completa, ci ha permesso di creare una class label più bilanciata: 67.80% di aziende attive e 33.20% fallite. Successivamente abbiamo creato l'attributo binario Failed il cui valore è 'Yes' se l'azienda è fallita (non attiva), 'No' altrimenti, indipendentemente dallo stato.

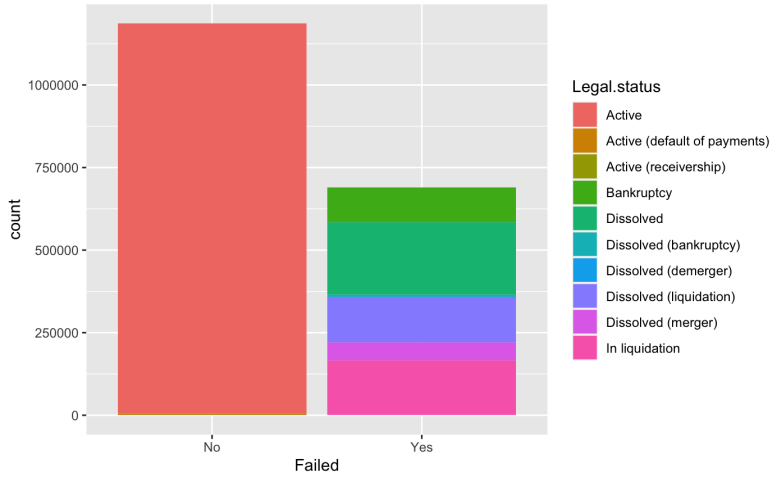



Figure 1.1: Numero di aziende attive e fallite, in base allo stato legale in cui si trova l'impresa.

## 1.3 Settore ATECO

Successivamente abbiamo creato la variabile Ateco.Sector.Name, la quale, in funzione della ripartizione delle attività economiche ATECO 2007 redatta dall'ISTAT, classifica il settore di ogni azienda (Classificazione ISTAT) con una determinata lettera. In particolare, ogni azienda viene

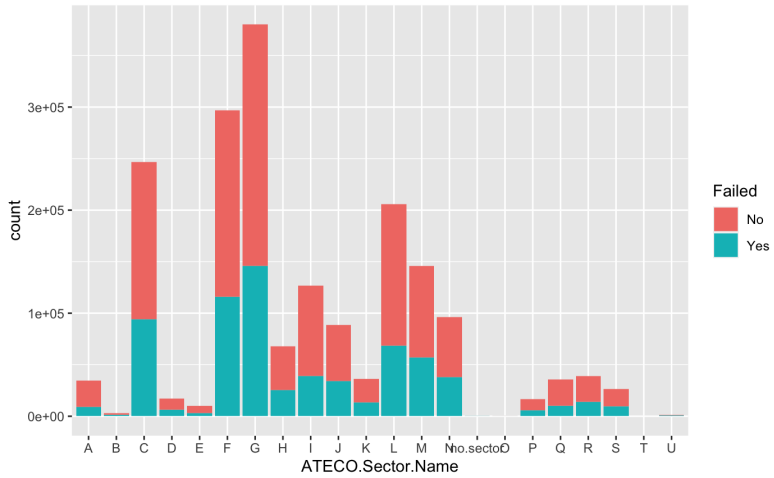



Figure 1.2: Ripartizione delle aziende in base al settore, distinguendo tra quelle attive e fallite.

identificata dall'ATECO.2007code, un codice numerico di 6 cifre che identifica l'attività caratteristica di quell'azienda, per un totale di 1678 codici differenti. Il settore viene identificato dalle prime due cifre con delle lettere dalla 'A' alla 'U'. La figura 1.2 mostra il numero di aziende per settore distinguendone lo stato operativo.

## 1.4 Data Preprocess

Il dataset assoluto che abbiamo creato riporta gli indicatori di una azienda in uno specifico anno. Al dataset sono state aggiunte le variabili Failed e ATECO.Sector.Name.
Per poter rispondere alle domande A, A.ctd e B dobbiamo necessariamente definire la variabile Age e Size, quindi:

- Creazione della variabile (i.e. Age non è presente nel dataset)
- Gestione dei missing values
- Analisi univariata delle anomalie: metodo IQR
- Eventuali trasformazioni della variabile (i.e. log o square root trasformation)

La variabile discreta **Age** rappresenta l'età che l'azienda possiede in un determinato anno ed è stata definita come la differenza tra Year e Incorporation.year. Lo 0.035% dei dati di quest'ultima è mancante. Anche se la quantità dei missing values è irrisoria abbiamo deciso di sostituire i valori mancanti con il valore mediano di Incorporation.year considerando le aziende dello stesso settore. In questo modo abbiamo ottenuto un risultato più preciso piuttosto che considerare la mediana su tutta la distribuzione. Successivamente abbiamo creato la variabile Age e abbiamo cancellato alcune aziende che riportavano un'età negativa. I boxplot hanno confermato la presenza di valori anomali esternamente ai whiskers. Le anomalie rappresentano il 18.34% della distribuzione e riguardano tutte le aziende con un'età strettamente superiore a 35 anni. Ovviamente non abbiamo eliminato le anomalie, altrimenti avremmo perso le informazioni di tutte le aziende più vecchie. Tutte le informazioni per rispondere alle domande che riguardano la variabile Age sono state esportate nel file *aida.age.RData*. Il sub-dataset pulito contiene 5,192,149 record e 7 features: Age, Legal.form, ATECO.Sector.Name, Region, Year, Failed e Company.Name.

| | Age | Size |
|---|---|---|
| Min | 0.00 | 0.00 |
| 1st Qu | 3.00 | 62.23 |
| Median | 7.00 | 209.06 |
| Mean | 11.51 | 433.15 |
| 3rd Qu | 16.00 | 597.20 |
| Max | 118.00 | 2415.30 |

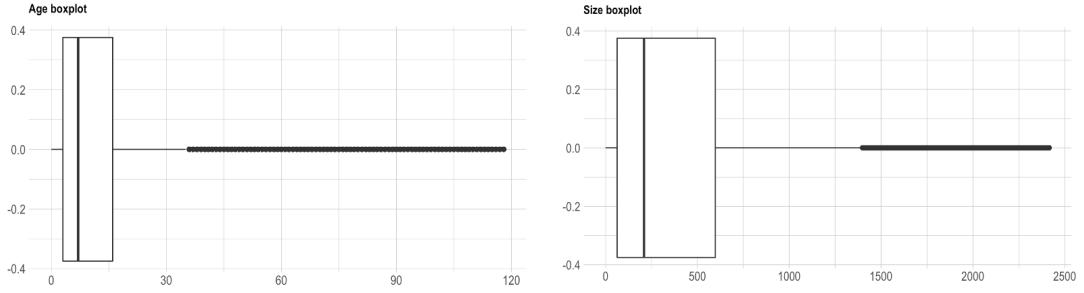



Figure 1.3: Boxplot di Age e Size. I valori anomali rappresentano rispettivamente le aziende più vecchie e quelle con un attivo maggiore.

La **dimensione** di una azienda è stata espressa in funzione del totale delle attività riportate in bilancio. Tale informazione è contenuta nella variabile continua Total.assetsth, pertanto non necessita di essere creata. Dopo aver rimosso il 10.7% dei valori mancanti abbiamo notato che l'87.06% della distribuzione era totalmente concentrata verso il basso. Quindi, abbiamo rimosso il 12.9% delle anomalie così da riportare le statistiche a valori più corretti. Come per Age, abbiamo esportato le informazioni nel file *aida.size.RData*. Il file processato contiene 4,372,481 record e 7 variabili: Total.assetsth, Legal.form, Region, Year, Failed, ATECO.Sector.Name e Company.name

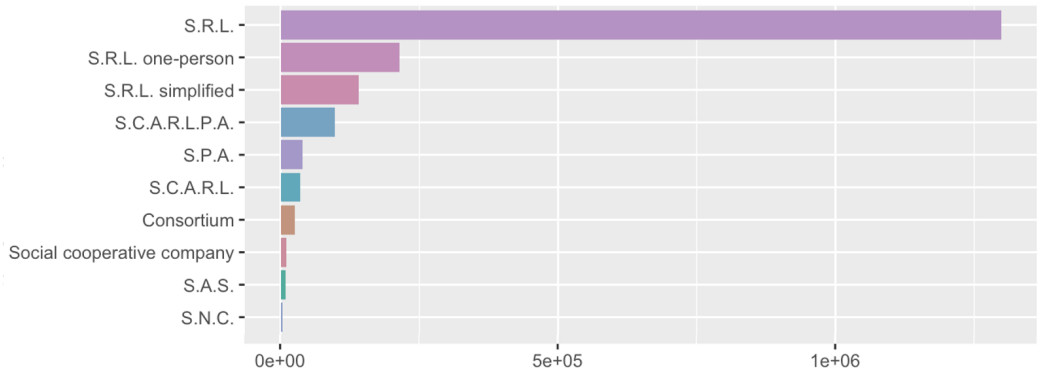



Figure 1.4: Top-10 Legal forms

Inoltre, abbiamo cancellato tutte le aziende per cui non era specificata la forma societaria, le quali costituiscono lo 0.02% del dataset. La figura 1.4 mostra il numero di aziende presenti per forma societaria. Abbiamo considerato le 10 più frequenti a cui ci atterremo quando dovremmo fare differenti tipi di analisi in base alla legal form.

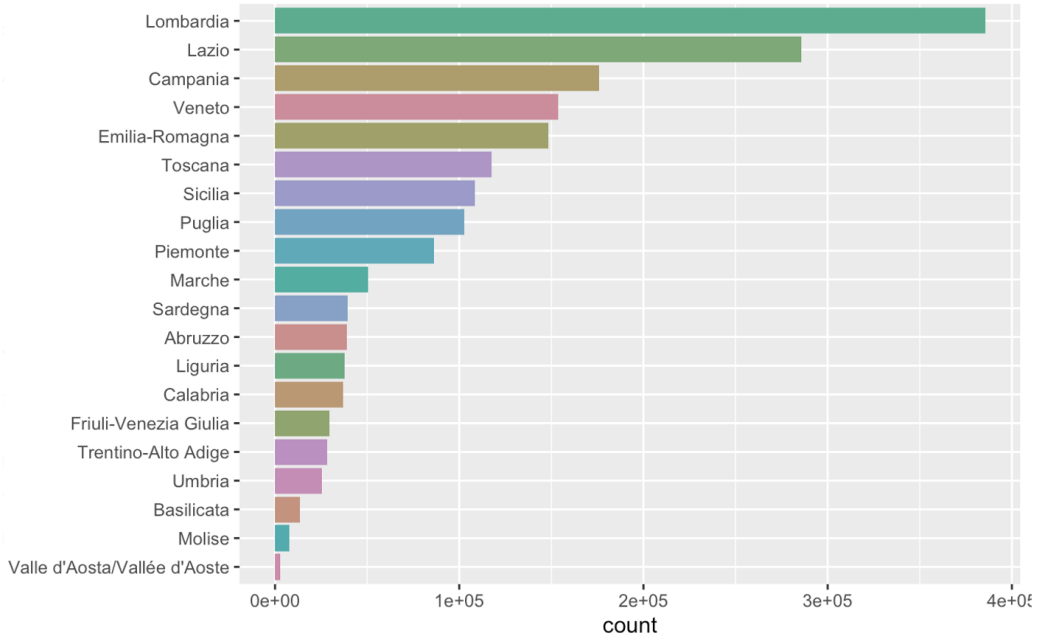



Figure 1.5: Numero di aziende per ogni regione

# 2 Domanda A

In questa sezione descriviamo le scelte e le analisi relativamente alla prima domanda. Ovvero valutare se: la distribuzione dell'età delle aziende e della loro dimensione cambia in maniera statisticamente significativa tra le aziende fallite e quelle attive in un determinato anno. Inoltre, confrontare le stesse distribuzioni al variare dei settori (ATECO) o della forma societaria (i.e. S.R.L., S.P.A., etc.).

## 2.1 Age: Analisi e Risultati Ottenuti

### 2.1.1 Distribuzione delle Età delle Aziende Attive e Fallite

Per confrontare la distribuzione dell'età delle aziende attive e fallite in un determinato anno, abbiamo utilizzato il t-test. Anche se l'età non segue una distribuzione normale abbiamo potuto applicare il t-test in quanto il numero di osservazioni è molto ampio. Non è stato possibile utilizzare il test di Kolmogorov-Smirnov perché la distribuzione di Age non è continua. Per il test abbiamo utilizzato un livello di significativà $\alpha = 0.05$, rigettando l'ipotesi nulla se e solo se il p-value è inferiore a $\alpha$.

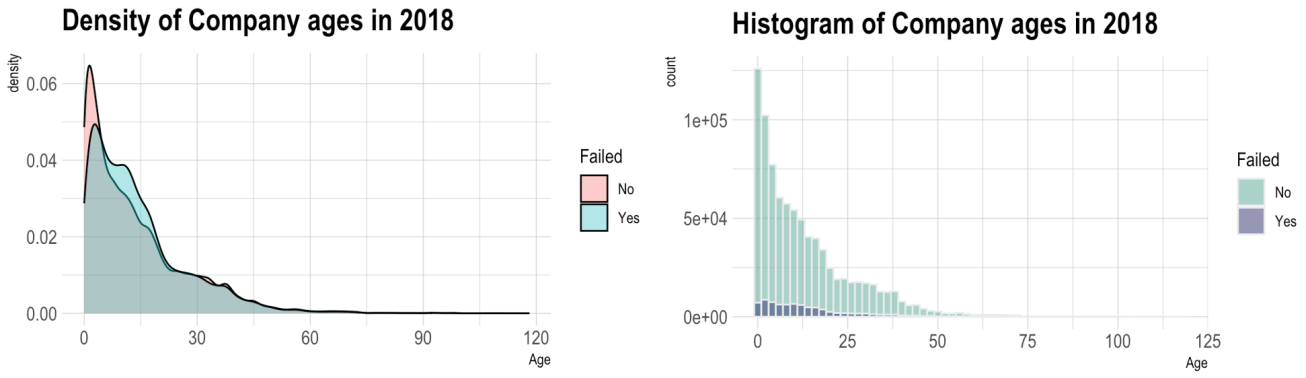



Figure 2.1: Densità e frequenza della distribuzione di Age

Abbiamo analizzato l'età di tutte le aziende nell'anno 2018. Come possiamo osservare le aziende attive hanno una densità maggiore per valori dell'età più vicini allo 0, mentre quelle fallite hanno una concentrazione inferiore inizialmente. Il t-test confronta rispettivamente l'età media delle aziende attive con quella delle aziende fallite.

| Test | Ipotesi * | p-value |
|---|---|---|
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A \neq \mu_F$ | <2.2e-16 |
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A < \mu_F$ | <2.2e-16 |
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A > \mu_F$ | $= 1$ |

Table 2.1: (*) $\mu_A$ = media reale dell'età delle aziende attive nel 2018, $\mu_F$ = media reale dell'età delle aziende fallite nel 2018

Come possiamo leggere, l'ipotesi nulla dei primi due test viene scartata a favore dell'ipotesi alternativa unilaterale sinistra, ciò significa che nel 2018, le aziende attive sono mediamente più giovani di quelle fallite.

### 2.1.2 Analisi al Variare del Settore e della Forma Societaria

Adesso valutiamo come cambia la distribuzione dell'età tra aziende attive e fallite, al variare della forma societaria e del settore a cui appartiene l'azienda. In particolare il test viene eseguito se abbiamo in entrambe le distribuzioni un numero di osservazioni minimo superiore a 100. Questo perché la ripartizione delle aziende fallite e attive in alcuni settori e/o company form potrebbe essere insufficiente, come riportato nelle figure 1.2 e 1.4.
Da un'analisi dei risultati possiamo notare che non è sempre vero che le aziende attive sono mediamente più giovani di quelle fallite nel 2018 se consideriamo i test per i singoli settori e le singole forme societarie.
Ad esempio, considerando la Forma Societaria S.C.A.R.L.P.A, risulta che le aziende attive hanno un età maggiore di quelle fallite. Per quanto riguarda il settore, considerando le aziende di tipo Q, si riscontra lo stesso fenomeno. In entrambi i casi, la differenza tra le medie è stata verificate con t-test ottenendo un p-value vicino a 0. Dunque, l'ipotesi nulla è stata scartata a favore dell'ipotesi alternativa unilaterale destra.

## 2.2 Size: Analisi e Risultati Ottenuti

### 2.2.1 Distribuzione delle Dimensioni delle Aziende Attive e Fallite

Il confronto della distribuzione della dimensione delle aziende attive e fallite in un certo anno è stato realizzato sia con il t-test che con il ks-test, in quanto la distribuzione è continua. Come per l'età, anche la distribuzione della dimensione non è normale ma avendo un elevato numero di osservazioni non ci poniamo il problema. Da sottolineare il fatto che se avessimo considerato il numero di dipendenti per misurare la dimensione di una azienda, in quel caso avremmo avuto una distribuzione normale. Le considerazioni sul livello di significatività sono le stesse di Age.

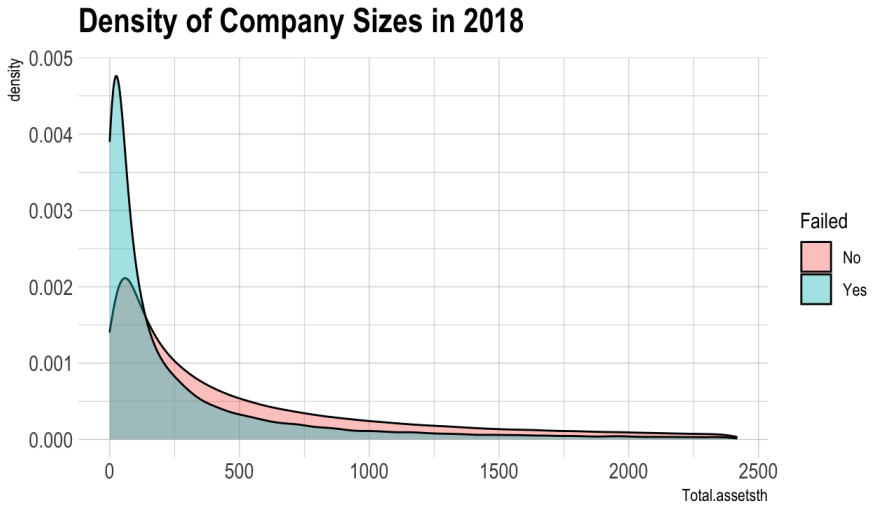



Figure 2.2: Densità e frequenza della distribuzione di Age

Dalla densità possiamo notare che gran parte delle aziende fallite ha una dimensione molto bassa. Il t-test respinge l'ipotesi nulla a favore dell'ipotesi alternativa unilaterale destra, affermando che, nel 2018, le aziende attive sono mediamente più grandi (in termini di attivo) di quelle fallite. Sommando questo risultato a quello di Age, possiamo affermare che nel 2018 le aziende attive sono mediamente più grandi e più giovani di quelle fallite

| **Test** | **Ipotesi *** | **p-value** |
|---|---|---|
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A \neq \mu_F$ | <2.2e-16 |
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A < \mu_F$ | $= 1$ |
| t-test | $H_0$: $\mu_A = \mu_F$<br>$H_1$: $\mu_A > \mu_F$ | <2.2e-16 |
| ks-test | $H_0$: i campioni provengono dalla stessa distribuzione<br>$H_1$: i campioni provengono da distribuzioni diverse | <2.2e-16 |

Table 2.2: (*) $\mu_A$ = media reale della dimensione delle aziende attive nel 2018, $\mu_F$ = media reale della dimensione delle aziende fallite nel 2018

### 2.2.2 Analisi di al Variare del Settore e della Forma Societaria

Come fatto per l'età di un'azienda, si ripete il processo di verifica delle distribuzioni della dimensione al variare della forma societaria e del settore. Tuttavia, in entrambi i casi, non si riscontrano variazioni del trend e quindi anche al variare di settore e forma societaria, la dimensione delle aziende attive è maggiore della dimensione di quelle fallite.

# 3 Domanda A.ctd

In questa sezione viene analizzata la seconda domanda, la quale rappresenta un approfondimento della precedente. In particolare, l'obiettivo è quello di valutare se la distribuzione dell'età e della dimensione delle aziende cambia in modo significativo tra le aziende fallite valutate in un certo arco temporale. Successivamente, osserviamo se al cambiare della forma societaria e della regione in cui l'azienda opera, la distribuzione varia.

## 3.1 Age: Analisi e Risultati Ottenuti

A differenza della prima domanda, consideriamo le sole aziende fallite testandone la distribuzione in un certo arco temporale. L'idea è stata quella di eseguire un t-test considerando l'età o la dimensione delle aziende fallite in un arco temporale di 2 anni. Successivamente, in base ai risultati ottenuti, estendere l'arco temporale su più anni eseguendo dei test multipli con ANOVA e Friedman, quest'ultimo nel caso in cui non si abbia la normalità delle distribuzioni.

### 3.1.1 Distribuzione delle Età delle Aziende Fallite su 2 anni

Dopo aver fatto diverse prove abbiamo scelto come campioni l'età di tutte le aziende fallite nel 2010 e nel 2011. Abbiamo utilizzato il t-test anche se le osservazioni non hanno una distribuzione normale, dato che il numero di osservazioni è molto elevato. Anche in questo caso il livello di significatività è $\alpha = 0.05$ e scartiamo $H_0$ nel caso in cui il p-value sia inferiore di $\alpha$.

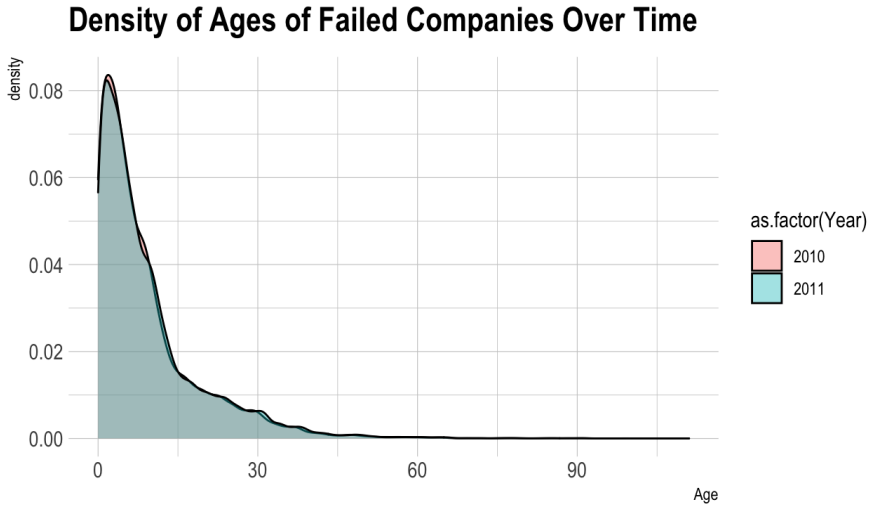



Figure 3.1: Densità delle aziende fallite nel 2010 e 2011

| Test | Ipotesi * | p-value |
|---|---|---|
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} \neq \mu_{F2}$ | <2.2e-16 |
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} < \mu_{F2}$ | <2.2e-16 |
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} > \mu_{F2}$ | = 1 |

Table 3.1: (*) $\mu_{F1}$ = media reale dell'età delle aziende fallite nel 2010, $\mu_{F2}$ = media reale dell'età delle aziende fallite nel 2011

Come possiamo osservare le densità delle distribuzioni sono molto simili. Il test respinge $H_0$ nelle prime due prove a favore dell'ipotesi alternativa unilaterale sinistra. Tale risultato afferma che le aziende fallite nel 2010 sono mediamente più giovani di quelle fallite nel 2011.

### 3.1.2 Analisi al Variare della Regione e della Forma Societaria

Abbiamo verificato se, anche al variare della forma societaria e della regione, le aziende fallite nel 2010 erano più giovani rispetto a quelle fallite nell'anno successivo. Considerando i campioni significativi, si riscontra che solamente per la tipologia S.N.C. si ha che l'età delle aziende fallite nel 2010 è maggiore dell'età delle aziende fallite nel 2011. Al variare della regione, invece, il trend rimane invariato.

### 3.1.3 Analisi su 5 Anni: ANOVA & Friedman Test

A questo punto abbiamo provato ad estendere l'arco temporale considerando oltre al 2010, i 4 anni successivi, così da avere 5 distribuzioni dell'età delle aziende fallite dal 2010 al 2014. Non abbiamo potuto utilizzare ANOVA perchè lo Shapiro test ha confermato la non normalità di tutte le distribuzioni. Di conseguenza è stato eseguito il test non parametrico di Friedman. Per eseguire il test tutte le distribuzioni devono avere lo stesso numero di osservazioni. Di conseguenza abbiamo considerato il minor numero $m$ di aziende fallite negli anni selezionati e per ogni distribuzione selezionato un campione casuale di lunghezza $m$.

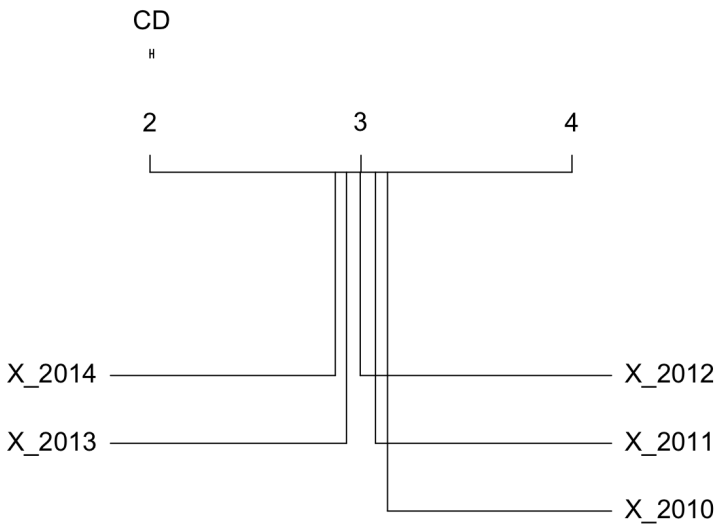



Figure 3.2: Critical difference plot per l'età

Il test di Friedman ha restituito un p-value $< \alpha$. Di conseguenza le mediane delle distribuzioni delle aziende fallite nei 5 anni selezionati non sono tutte simili. Quindi, abbiamo eseguito il post-hoc test di Nemenyi per individuare quali sono le coppie di distribuzioni differenti. I risultati del test mostrano che tutte le coppie di campioni sono tra loro diverse.

## 3.2 Size: Analisi e Risultati Ottenuti

### 3.2.1 Distribuzione delle Dimensioni delle Aziende Fallite su 2 anni

Abbiamo effettuato il ks-test e il t-test, considerando sempre le aziende fallite nel 2010 e 2011. Il livello di significatività con cui confrontare il p-value ritornato dai test è sempre pari a $\alpha = 0.05$.

I due test restituiscono risultati discordanti. Il ks-test ritorna un p-value $< \alpha$, respingendo

| **Test** | **Ipotesi *** | **p-value** |
|---|---|---|
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} \neq \mu_{F2}$ | 0.2002 |
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} < \mu_{F2}$ | 0.1001 |
| t-test | $H_0$: $\mu_{F1} = \mu_{F2}$<br>$H_1$: $\mu_{F1} > \mu_{F2}$ | 0.8999 |
| ks-test | $H_0$: i campioni provengono dalla stessa distribuzione<br>$H_1$: i campioni provengono da distribuzioni diverse | 0.04056 |

Table 3.2: (*) $\mu_{F1}$ = media reale della dimensione delle aziende fallite nel 2010, $\mu_{F2}$ = media reale della dimensione delle aziende fallite nel 2012

l'ipotesi nulla a favore dell'ipotesi alternativa bilaterale, affermando la diversità delle distribuzioni. Il t-test, invece, in tutti e tre i test, restituisce sempre un p-value superiore a $\alpha$. Questo non ci permette di scartare l'ipotesi nulla nè di affermare che le medie delle distribuzioni sono mediamente uguali.
Una possibile causa di questi risultati contrastanti potrebbe essere dovuta dal modo in cui i test confrontano le distribuzioni; il t-test utilizza la media della popolazione, mentre il test di Kolmogorov-Smirnov si basa sulla distanza tra le funzioni di ripartizione empiriche dei due campioni.

### 3.2.2 Analisi al Variare della Regione e della Forma Societaria

Successivamente abbiamo verificato se al variare della regione e della forma societaria, le dimensioni delle aziende fallite nei 2 anni erano sempre simili o esistevano dei casi in cui poteva non valere questa relazione. Tra i risultati più significativi abbiamo che per la tipologia S.R.L., il p-value restituito dal test permette di scartare l'ipotesi nulla a favore dell'ipotesi alternativa unilaterale sinistra. Quindi, il test evidenzia come la dimensione delle aziende S.R.L. fallite nel 2010 è minore alla dimensione della aziende S.R.L. fallite nel 2011. Al variare delle Regione, tutti i test effettuati non permettono di scartare l'ipotesi nulla a favore di quelle alternative. Tuttavia, guardando al valore del p-value si può comunque capire qual'è la relazione, anche se non significativa, per le distribuzioni considerate. Concludendo, dall'analisi delle Regioni, non si hanno cambiamenti significativi rispetto a quelli osservati precedentemente 3.2.1.

### 3.2.3 Analisi su 5 Anni: ANOVA & Friedman Test

Visto il risultato statistico del t-test sui campioni di 2 anni, abbiamo provato ad estendere l'arco temporale considerando i 4 anni successivi: quindi dal 2010 al 2014. Non abbiamo potuto utilizzare ANOVA in quando il test di Shapiro ha confermato la non normalità dei campioni. Anche in questo caso abbiamo utilizzato il test multiplo non parametrico di Friedman. I campioni, come in Age, sono stati selezionati casualmente considerando la lunghezza minima delle 5 distribuzioni.

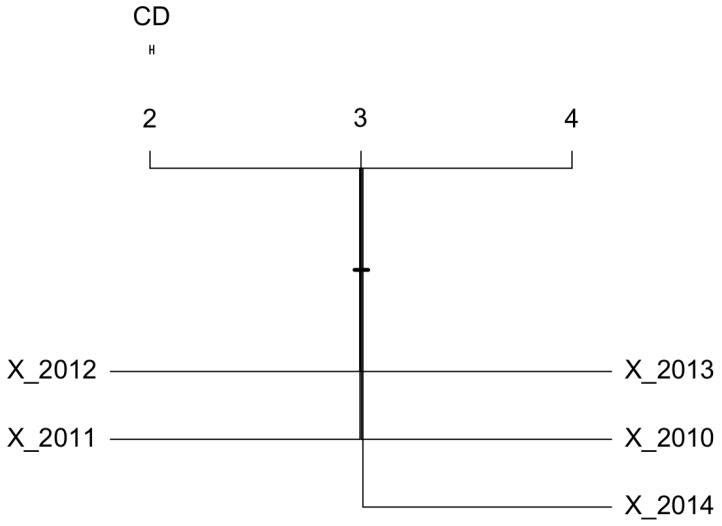



Figure 3.3: Critical difference plot per la dimensione

Il test di Friedman ritorna un p-value = $0.147 > \alpha$, questo ci porta a non scartare l'ipotesi nulla, esattamente come nel t-test. Non essendo ancora arrivati ad una conclusione certa, abbiamo eseguito il post-hoc test di Nemenyi.
Il test afferma che per tutte le coppie di distribuzioni non si può escludere la possibilità che siano uguali: cioè che le dimensioni delle aziende fallite dal 2010 al 2014 sono simili.

# 4 Domanda B

La terza domanda richiede di analizzare in un determinato anno, la distribuzione dei fallimenti conoscendo l'età o la dimensione delle aziende. Successivamente approfondire l'analisi valutando se tale distribuzione cambia in funzione della forma societaria, del settore o della regione in cui opera l'azienda.

Questa domanda si differisce dalle precedenti, in quanto ci viene richiesto di calcolare la probabilità in un certo anno, che si verifichi un fallimento conoscendo l'età o la dimensione dell'azienda:

$$P(Failed = Yes|Age/Size = val) = \frac{P(Failed = Yes, Age/Size = val)}{P(Age/Size = val)}$$

Ciò si riconduce a calcolare il seguente rapporto:

$$P(Failed = Yes|Age/Size = val) = \frac{\#\ Aziende\ fallite\ tc\ Age/Size{=}val}{\#\ Aziende\ tc\ Age/Size{=}val}$$

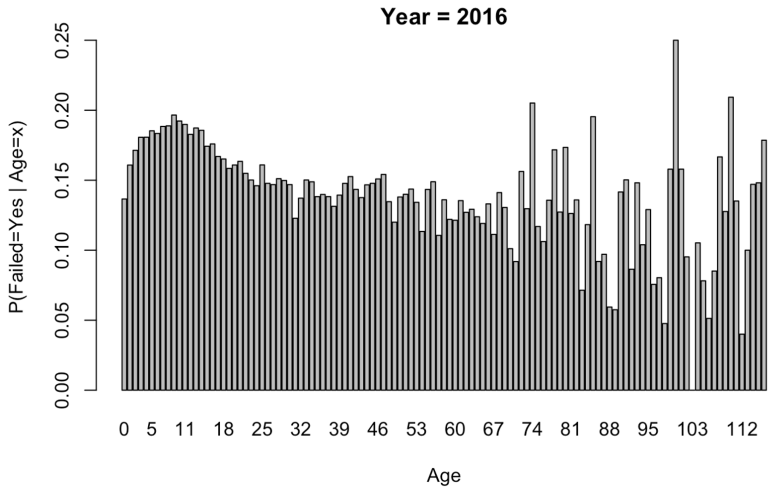
Quando l'analisi viene approfondita distinguendo i settori, le regiorni e le forme societarie, il calcolo della probabilità condizionata deve prendere in considerazione anche tale informazione (i.e. $P(Failed = Yes|Age/Size = val, Region = reg)$).

## 4.1 Age: Analisi e Risultati Ottenuti

### 4.1.1 Distribuzione Condizionata delle Età Aziendali

Abbiamo scelto come anno di riferimento il 2016 in quanto contiene una buona quantità di dati ad ogni età. A differenza della variabile Size, in questo caso non abbiamo discretizzato ulteriormente i dati in bins, in quanto i valori dell'età vanno da un minimo di 0 a un massimo di 116 anni nel 2016. Per ogni età abbiamo calcolato la distribuzione condizionata come segue:

$$P(Failed = Yes|Age = val) = \frac{\#\ Aziende\ fallite\ tc\ Age{=}val}{\#\ Aziende\ tc\ Age{=}val}$$




| | not.failed | failed | total | p.failed.knowing.age |
|---|---|---|---|---|
| 0 | 68432 | 10826 | 79258 | 0.13659189 |
| 1 | 61368 | 11761 | 73129 | 0.16082539 |
| 2 | 51641 | 10680 | 62321 | 0.17137081 |
| 3 | 42701 | 9416 | 52117 | 0.18067041 |
| 4 | 36133 | 7972 | 44105 | 0.18075048 |
| 5 | 35093 | 7982 | 43075 | 0.18530470 |
| 6 | 35510 | 7976 | 43486 | 0.18341535 |
| 7 | 30517 | 7085 | 37602 | 0.18842083 |
| 8 | 31289 | 7284 | 38573 | 0.18883675 |
| 9 | 30327 | 7419 | 37746 | 0.19655063 |
| 10 | 28686 | 6829 | 35515 | 0.19228495 |
| 11 | 26658 | 6248 | 32906 | 0.18987419 |
| 12 | 24503 | 5479 | 29982 | 0.18274298 |
| 13 | 20424 | 4708 | 25132 | 0.18733089 |

Figure 4.1: A sinistra la distribuzione dei fallimenti conoscendo l'età delle aziende del 2016. A destra una parte della tabella delle probabilità condizionate per ogni valore di age.

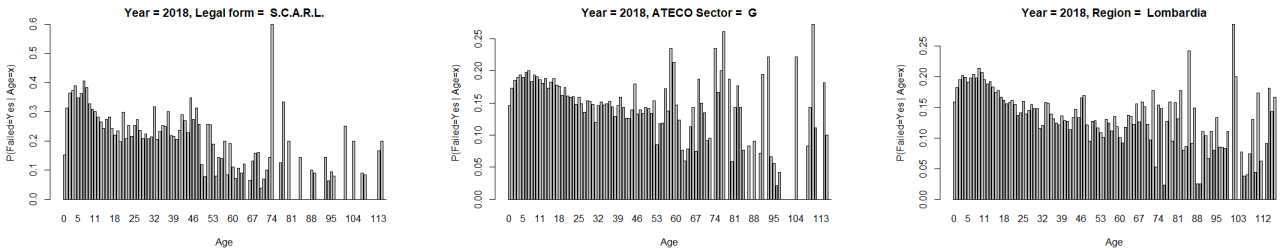
Come possiamo osservare dai risultati in Figura 4.1, all'aumentare dell'età la probabilità di fallimento tende mediamente a diminuire. I picchi isolati che osserviamo a partire da un'età superiore a 70 anni sono causati da una scarsa quantità di aziende presenti con quell'età. Ad esempio la probabilità massima la abbiamo con un età pari a 100 anni, dove abbiamo soltanto 12 aziende nel 2016, di cui 3 sono fallite. Se paradossalmente ad una certa età avessimo soltanto aziende fallite, la probabilità condizionata varrebbe 1.

### 4.1.2 Analisi al Variare del Settore, Forma Societaria e Regione

Al variare della forma societaria, dei settori e delle regioni abbiamo deciso di effettuare dei test grafici per confrontare se l'andamento della probabilità con quello generale. Non si sono effettuati test statistici per verificare se la probabilità di ogni bin varia rispetto alla distribuzione di probabilità generale dato che la variabilità in ogni intervallo è grande (Ad esempio per Age $>$ 70 abbiamo probabilità meno significative dovute a mancanza di osservazioni).

Si è proseguito con l'analisi grafica dell'andamento delle probabilità, guardando solo ai casi con un numero minimo di osservazioni.

In Figura 4.2 si sono presi in analisi tre casi: S.C.A.R.L. per Legal Form, G per Sector e Lombardia per la regione. Si osserva che l'andamento di probabilità segue quello osservato per il caso generale. Nel caso di S.C.A.R.L. si ha inoltre che la probabilità di fallimento è superiore per valori di Age bassi rispetto al caso generale.




(a) (b) (c)

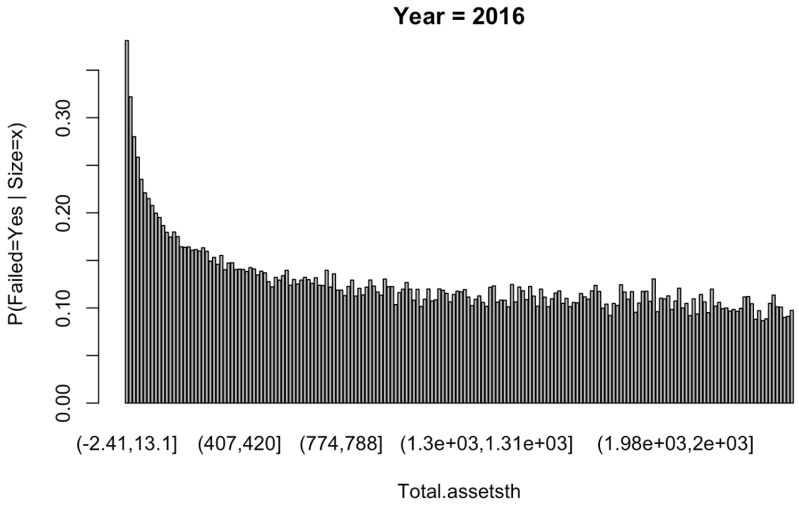
Figure 4.2: Analisi al variare di Legal Form, Sector e Region

## 4.2 Size: Analisi e Risultati Ottenuti

### 4.2.1 Distribuzione Condizionata delle Dimensioni Aziendali

Abbiamo discretizzato la variabile Size in un determinato numero di intervalli di ugual ampiezza, utilizzando il metodo di Freedman-Diaconis. Le dimensioni delle aziende nel 2016 vanno da un minimo di 0.001 ad un massimo di 2415.224. I bins ottenuti con Freedman sono 184 ed hanno un ampiezza di 13.15341. La distribuzione condizionata viene calcolata come segue:

$$P(Failed = Yes|Size.group = val) = \frac{\#\ Aziende\ fallite\ che\ sono\ nello\ stesso\ intervallo}{\#\ Aziende\ che\ sono\ nello\ stesso\ intervallo}$$




| | not.failed | failed | total | p.failed.knowing.size |
|---|---|---|---|---|
| (-2.41,13.1] | 32686 | 20119 | 52805 | 0.38100559 |
| (13.1,26.3] | 24429 | 11597 | 36026 | 0.32190640 |
| (26.3,39.4] | 22598 | 8787 | 31385 | 0.27997451 |
| (39.4,52.5] | 21715 | 7565 | 29280 | 0.25836749 |
| (52.5,65.6] | 19914 | 6126 | 26040 | 0.23525346 |
| (65.6,78.8] | 18637 | 5285 | 23922 | 0.22092634 |
| (78.8,91.9] | 17441 | 4779 | 22220 | 0.21507651 |
| (91.9,105] | 16488 | 4322 | 20810 | 0.20768861 |
| (105,118] | 15617 | 3893 | 19510 | 0.19953870 |
| (118,131] | 14610 | 3544 | 18154 | 0.19521868 |
| (131,144] | 13662 | 3133 | 16795 | 0.18654361 |
| (144,158] | 12921 | 2829 | 15750 | 0.17961905 |
| (158,171] | 12302 | 2602 | 14904 | 0.17458400 |
| (171,184] | 11395 | 2499 | 13894 | 0.17986181 |
| (184,197] | 10990 | 2331 | 13321 | 0.17498686 |
| (197,210] | 10634 | 2092 | 12726 | 0.16438787 |
| (210,223] | 10103 | 1978 | 12081 | 0.16372817 |
| (223,236] | 9563 | 1877 | 11440 | 0.16407343 |
| (236,249] | 9220 | 1767 | 10987 | 0.16082643 |
| (249,263] | 8834 | 1699 | 10533 | 0.16130257 |
| (263,276] | 8605 | 1639 | 10244 | 0.15999610 |
| (276,289] | 8115 | 1582 | 9697 | 0.16314324 |
| (289,302] | 8018 | 1520 | 9538 | 0.15936255 |
| (302,315] | 7535 | 1324 | 8859 | 0.14945253 |
| (315,328] | 7174 | 1296 | 8470 | 0.15301063 |
| (328,341] | 7020 | 1200 | 8220 | 0.14598540 |

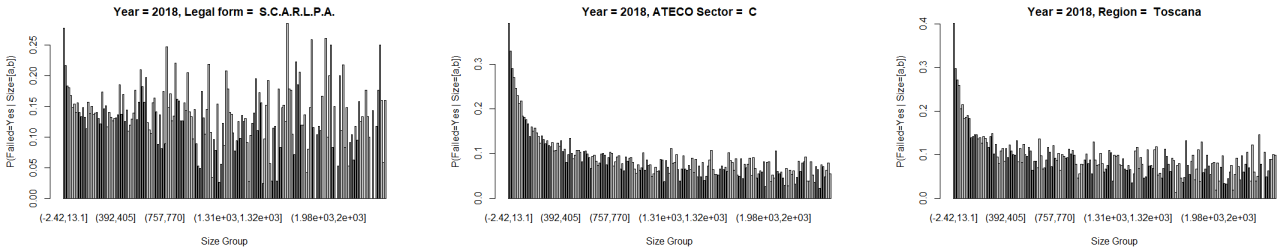
Figure 4.3: A sinistra la distribuzione dei fallimenti conoscendo le dimensioni delle aziende del 2016. A destra una parte della tabella delle probabilità condizionate per ogni intervallo di Size.

Come per Age, anche in questo caso abbiamo un trend decrescente. All'aumentare della dimensione abbiamo una minor probabilità di fallimento, mentre aziende con un attivo più basso hanno una probabiltà decisamente superiore di fallimento. Ad esempio, con un valore dell'attivo inferiore a 100 abbiamo una probabilità di fallimento che va dal 38.10% al 19.96% circa.

### 4.2.2 Analisi al Variare del Settore, Forma Societaria e Regione

Anche per la dimensione abbiamo è effettuato un'analisi grafica della probabilità al variare del Settore, Forma Societaria e Regione.

In Figura 4.4 si sono analizzati i 3 seguenti casi: S.C.A.R.L.P.A per Legal Form, C per Sector e Toscana per la regione. È possibile osservare che per Legal Form S.C.A.R.L.P.A. l'andamento di probabilità non segue la tendenza della distribuzione generale. Mentre per il settore C e la regione Toscana, l'andamento sembra seguire quello osservato per la distribuzione generale.




(a) (b) (c)

Figure 4.4: Analisi al variare di di Legal Form, Sector e Region

# 5 Domanda C

In questa sezione viene descritto il procedimento e i risultati ottenuti per rispondere alla domanda C. Per creare uno scoring model che restituisce la probabilità che un'azienda sia fallita e un rating model che classifica l'azienda Active/Failed, si è utilizzato la Regressione Logistica con e senza regolarizzazione.

In aggiunta ai due modelli parametrici, si è realizzato un modello Random Forest in modo da poter effettuare un confronto diretto con le performance della Regressione Logistica.

## 5.1 Preprocessing

Il dataset utilizzato in questa fase è composto da tutti gli indicatori finanziari presenti e quindi privato degli attributi categorici. Inoltre, si è rimosso tutti quei record aventi valori mancanti. Per decidere quali anni usare si è guardato alla distribuzione della classe target in modo da ottenere un dateset bilanciato. Utilizzando i soli dati relativi agli anni 2014 e 2015 si ottiene un dataset con 20.565 record con classe 0 e 21.084 record con classe 1. Infine, si è partizionato il dataset in training set (70%) e test set (30%). In aggiunta al test set relativo agli anni 2014 e 2015, si è creato un test set relativo all'anno 2013 per valutare il comportamento del modello con record di altri anni.

## 5.2 Regressione Logistica

In questa sezione viene descritto il procedimento e i risultati ottenuti con Regressione Logistica senza regolarizzazione. La selezione delle variabili da utilizzare per il modello finale si compone di due fasi:

1. *Rimozione delle variabili che causano multicollinearità*: variabili con alta correlazione rendono i coefficienti della regressione instabili e aumentano lo standard error. Si sono individuate le coppie di variabili con correlazione maggiore di 0.8. Successivamente, per decidere quali variabile rimuovere, si è utilizzato il valore di Variance Inflaction Factor ottenuto da una regressione logistica sull'intero dataset. In Tabella 5.1 sono riportati i confronti effettuati, la variabile rimossa è quella con valore di VIF maggiore.

| $x_1$ | $vif_{x1}$ | $x_2$ | $vif_{x2}$ | $\rho$ |
|---|---|---|---|---|
| ROA | 5.38 | ROI | 4.63 | 0.859 |
| Number.of.employees | 25.53 | Net.financial.positiont | 522.70 h | 0.812 |
| Cash.Flowth | 29.20 | Profit.(loss)th | 6.86 | 0.902 |
| Net.financial.positionth | 522.70 | Net.working.capitalth | 267.95 | 0.916 |
| Liquidity.ratio | 2.99 | Current.ratio | 3.44 | 0.811 |

Table 5.1: Coppie di attributi correlati e VIF associato

2. *Backward step con AIC*: Utilizzando Akaike Informatio Criterion si cerca il miglior sottoinsieme di attributi in termini di fitting del modello e semplicità.

Seguendo la procedura descritta si è ottenuto un modello con le seguenti variabili:

```
Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          3.267e+01  2.284e+00  14.302  < 2e-16 ***
Incorporation.year                  -1.661e-02  1.139e-03 -14.584  < 2e-16 ***
`\\`Banks/turnover\\``               2.415e-03  9.630e-04   2.508 0.012153 *
Cost.of.debit                       -7.904e-03  3.495e-03  -2.262 0.023721 *
`\\`Current.liabilities/Tot.ass\\``  5.176e-01  6.026e-02   8.590  < 2e-16 ***
`\\`Debt/equity.ratio\\``            2.122e-03  8.922e-04   2.378 0.017385 *
`\\`EBITDA/Vendite\\``              -5.833e-03  1.624e-03  -3.593 0.000327 ***
`\\`Interest/Turnover\\``            1.635e-02  1.112e-02   1.470 0.141477
Liquidity.ratio                      4.887e-02  1.471e-02   3.322 0.000895 ***
Net.working.capitalth               -8.133e-06  2.927e-06  -2.779 0.005453 **
Number.of.employees                  5.174e-04  2.055e-04   2.518 0.011815 *
`\\`Profit.(loss)th\\``              4.530e-05  1.915e-05   2.365 0.018019 *
ROE                                 -3.197e-03  4.289e-04  -7.453 9.13e-14 ***
ROS                                 -6.968e-03  2.140e-03  -3.255 0.001132 **
Solvency.ratio                      -2.586e-03  7.447e-04  -3.473 0.000515 ***
`\\`Total.assets.turnover.(times)\\`` 2.015e-01  1.864e-02  10.810  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.1: Coefficienti Regressione Logistica

Per comprendere l'utilità di una variabile per la classificazione si sono osservati i coefficienti calcolati dal modello, in particolare, guardando al segno si può capire se tale variabile contribuisce a classificare un'azienda come Failed (segno posivito) oppure come Active (segno negativo).

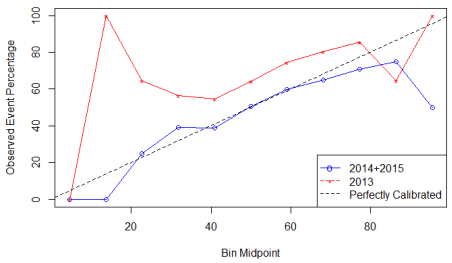
Tra le variabili più significative con segno positivo abbiamo Current.liabilities/Tot.ass con coefficiente 0.517, Total.assets.turnover.(times) con coefficiente 0.201 e l'intercept con valore 32.67. Il p-value per questi valori è molto basso e quindi si può rifiutare l'ipotesi che i coefficienti siano uguali a 0.

Successivamente si è proseguito con la valutazione delle performance sul test set nei vari anni (2013, 2014 e 2015). In Tabella 5.2 sono riportati i valori di Accuracy, F1, Precision e Recall per i due test set.

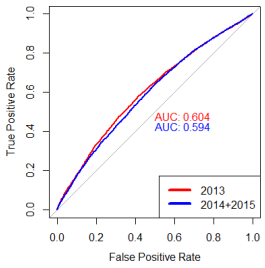
| Anni | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2013 | 59 | 67 | 72 | 62 |
| 2014+2015 | 57 | 57 | 57 | 57 |

Table 5.2: Performance al variare degli anni

In aggiunta, si sono valutati i grafici Calibration Plot e ROC Curve come misure di qualità per i due anni 5.2.




(a) Calibration Plot




(b) ROC Curve e AUC

Figure 5.2

Guardando al Calibration Plot, il modello sembra sovrastimare la quantità di aziende Failed per l'anno 2013, mentre per gli anni 2014+2015 si avvicina maggiormente ad un modello perfettamente calibrato.

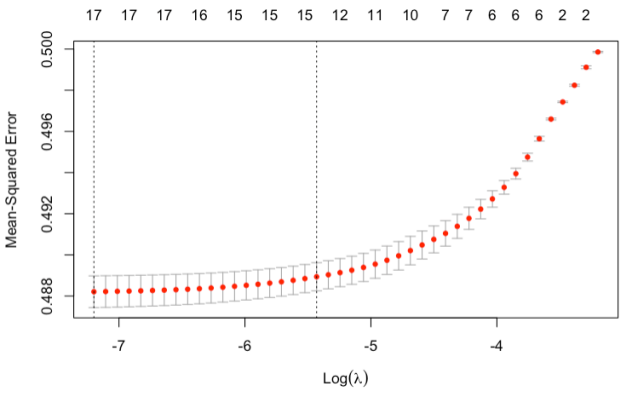
Tuttavia, guardando al grafico ROC Curve e al valore di AUC sembrerebbe che il modello valutato sull'anno 2013 abbia performance migliori.

### 5.2.1 Elasticnet Regularization

In questa sezione verrà descritto come abbiamo implementato, ed i risultati che abbiamo ottenuto regolarizzando il modello di regressione logistica utilizzando il metodo Elasticnet.

Per rendere possibile una comparazione dei risultati ottenuti da i due regressori, abbiamo utilizzato la stessa task descritta nella sezione 5.2 dell'elaborato, includendo quindi nel training set solo misurazioni relative agli anni 2014 e 2015, infine, abbiamo testato il classificatore utilizzando sia un test set relativo agli anni 2014 e 2015, che uno con osservazioni riguardanti esclusivamente il 2013.

Per allenare il modello abbiamo utilizzato la libreria *glmnet*, la quale utilizza la cross validation sul training set al fine di trovare il miglior valore di lambda in grado di regolarizzare il regressore.




(a) Mean-Squared error al variare di lambda

| | s1 |
|---|---|
| (Intercept) | 2.889524e+01 |
| Incorporation.year | -1.475165e-02 |
| Banks/turnover | 1.686993e-03 |
| Cost.of.debit | -8.476499e-03 |
| Current.liabilities/Tot.ass | 5.581523e-01 |
| Debt/EBITDA.ratio | . |
| Debt/equity.ratio | 3.310005e-03 |
| EBITDA/Vendite | -3.813605e-03 |
| EBITDAth | 1.665955e-06 |
| Interest/Turnover | 2.904879e-02 |
| Leverage | 2.269741e-04 |
| Liquidity.ratio | 4.003431e-02 |
| Net.working.capitalth | . |
| Number.of.employees | 4.590357e-04 |
| Profit.(loss)th | 6.639277e-06 |
| ROE | -2.273818e-03 |
| ROI | -1.700915e-03 |
| ROS | -8.245202e-03 |
| Solvency.ratio | -2.284175e-03 |
| Total.assets.turnover.(times) | 2.116039e-01 |
| Total.assetsth | . |

(b) Coefficienti regressione logistica con Elastinet

Figure 5.3

Il metodo *cv.glmnet* ritorna quindi la lista di tutti i classificatori allenati e, tra questi, vengono forniti:

- *lambda.min*, valore di lambda che fornisce il regressore che minimizza la *Mean-Squared Error*.
- *lambda.1se*, il valore di lambda che ritorna il modello più regolarizzato, tale che l'errore sia entro uno *standard error* dal minimo.

Al fine di comprendere l'importanza nella classificazione di ogni variabile, abbiamo valutato i coefficienti di ogni attributo (riportati nella figura 5.3), possiamo vedere come tra le variabili negative più significative abbiamo *Cost.of.debit*, *ROS* ed *EBITDA/Vendite*, ciò vuol dire che, all'aumentare del valore di queste variabili, e più è alta la probabilità che l'azienda non sia fallita. Tra le veriabili positive più significative, invece, abbiamo: *Profit.(loss)th*, *Number.of.employees* e *Current.liabilities/Tot.ass*, ciò vuol dire che, all'aumentare del valore di queste variabili, aumenterà anche la probabilità che l'azienda sia fallita.

| | Lambda | Measure | SE | Nonzero |
|---|---|---|---|---|
| min | 0.000748 | 0.4882 | 0.0007704 | 17 |
| 1se | 0.004378 | 0.4889 | 0.0006833 | 14 |

Table 5.3: Valori di lambda restituiti da *cv.glmnet*

Abbiamo valutato le performance per entrambi i valori di lambda ritornati dalla funzione e abbiamo optato per il modello con lambda.1se, anche se, come si può vedere da 5.4, le performance ottenute con Lambda.min sono più alte, se si considera la precision nel test set relativo al 2014-2015. Abbiamo optato per questa scelta in quanto, a seguito di ricerche[1][2], abbiamo constatato che è buona pratica utilizzare il modello regolarizzato (lambda.1se) in quanto il classificatore con valore di lambda uguale a lambda.min è più soggetto ad overfitting.

| | **Lambda.min** | | | | **Lambda.1se** | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | F1 | Precision | Recall | Accuracy | F1 | Precision | Recall |
| 2014-2015 | 57 | 58 | 58 | 58 | 57 | 58 | 57 | 58 |
| 2013 | 59 | 67 | 72 | 63 | 59 | 67 | 72 | 63 |

Table 5.4: Score del modelli al variare degli anni

Confrontando le tabelle 5.2 e 5.4, si può vedere come, tramite il regressore allenato utilizzando Elasticnet per regolarizzare il modello siamo riusciti ad alzare leggermente le performance.

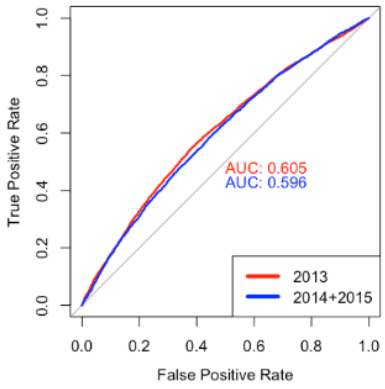



Figure 5.4: Confronto AUC tra test set 2013 e 2014-2015

Analogamente al regressore ottenuto senza regolarizzazione (come evidenziato in figura 5.4), abbiamo riscontrato un lieve innalzamento negli score ottenuti dai modelli nel classificare le

[1]Friedman, Hastie, and Tibshirani (2010), sezione "Selecting the tuning parameters"
[2]Krstajic, et al (2014)

aziende relative all'anno 2013, anche se questi sono stati allenati considerando gli anni 2014 e 2015.

## 5.3 Random Forest

In aggiunta ai modelli parametrici, si è realizzato un modello Random Forest con *n_tree*=50 e come numero di attributi per ogni albero $\sqrt{n}$ dove n è il numero totale di attributi.

Per avere delle misure di performance più significative si è utilizzato cross validation con 5 ripetizioni e 10 folds ognuna. In questo modo si sono ottenute 50 misurazioni di performance utilizzabili per un confronto statistico tra due o più modelli.

Le performance valutate sul test set sono riportate in Tabella 5.5.

| Anni | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| 2013 | 61 | 67 | 76 | 61 |
| 2014+2015 | 60 | 60 | 61 | 59 |

Table 5.5: Performance Random Forest al variare degli anni

## 5.4 Confronto delle performance dei modelli

In questa sezione si prosegue con un'analisi più approfondita delle performance raggiunte dai modelli calcolando degli intervalli di confidenza per le metriche di interesse, e successivamente, effettuando un test statistico tra le performance per valutare quale dei modelli ha raggiunto migliori risultati.

La metrica per la valutazione delle performance sulla quale si è effettuato le analisi è il valore di *AUC*.

Prima di tutto si sono raccolte le misurazioni di AUC sui fold della cross validation per i due modelli, in Figura 5.5 sono mostrati i risultati in un Boxplot.

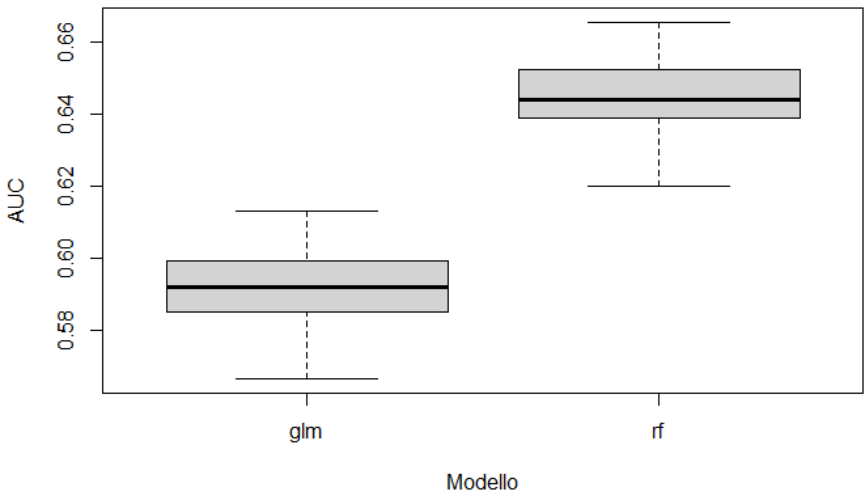



Figure 5.5: Boxplot dei valori di AUC della cross validation

Successivamente si è calcolato degli intervalli di confidenza per le due distribuzioni utilizzando i metodi *basic, norm, e bca* con livello di confidenza 95% su un bootstrap sample con 1000 ripetizioni. Gli intervalli di confidenza ottenuti sono riportati in Tabella 5.6.

| Metodo | glm | rf |
|---|---|---|
| basic | ( 0.5892, 0.5950 ) | ( 0.6423, 0.6479 ) |
| norm | ( 0.5891, 0.5950 ) | ( 0.6423, 0.6478 ) |
| bca | ( 0.5890, 0.5949 ) | ( 0.6421, 0.6477 ) |

Table 5.6: Intervalli di confidenza per AUC

Dalla Tabella 5.6 si può osservare che tutti i metodi utilizzati hanno restituito intervalli di confidenza molto simili.

Calcolati gli intervalli di confidenza per le due distribuzioni, si è proseguito con il confronto tra le medie delle due distribuzioni.

Il test di Shapiro per verificare la normalità ha restituito un p-value alto in entrambi i casi, quindi si può assumere che le due distribuzioni siano normali.

Successivamente, si è utilizzato F-test per confrontare le varianze delle due distribuzioni, il test ha restistuito un p-value alto e quindi si può assumere che le due distribuzioni abbiano la stessa varianza.

Infine, si sono eseguiti usa serie di test statistici con livello di confidenza 95% per confrontare le medie delle due distribuzioni 5.7.

| test | p-value |
|---|---|
| t-test (varianza uguale) | 2.2e-16 |
| Wilcoxon test | 2.2e-16 |
| boot t-test | 2.2e-16 |
| paired test | 2.2e-16 |

Table 5.7: Test statistici per confrontare le medie

È evidente come per tutti i test l'ipotesi nulla debba essere scartata e quindi si può assumere che le due distribuzioni hanno un valore medio differente.

## 5.5 Test statistici per il Rating

In quest'ultima sezione andremo a verificare la qualità del Rating di rischio fallimento che produce il modello di Regressione Logistica.

Fin dall'inizio si sono considerate le due classi di probabilità standard per identificare un'azienda come Attiva o Fallita.

Tuttavia, lo score di probabilità può suddiviso in decili per poi analizzare il numero di valori attesi e osservati delle due classi in ogni decile.

Il test del Rating di Hosmer-Lemeshow utilizza questo concetto di suddivisione della probabilità in gruppi per valutare la bontà di un fitting guardando ai valori attesi e osservati per le classi.

Si è dunque proseguito con il test del Rating usando 10 gruppi, i risultati sono riportati in Tabella 5.8.

| range | $\hat{y_0}$ | $\hat{y_1}$ | $y_0$ | $y_1$ |
|---|---|---|---|---|
| [0.156,0.413] | 1824.617 | 1102.383 | 1824 | 1103 |
| (0.413,0.444] | 1669.344 | 1257.656 | 1725 | 1202 |
| (0.444,0.465] | 1594.812 | 1331.188 | 1628 | 1298 |
| (0.465,0.483] | 1539.409 | 1387.591 | 1567 | 1360 |
| (0.483,0.5] | 1488.963 | 1438.037 | 1477 | 1450 |
| (0.5,0.518] | 1437.726 | 1488.274 | 1412 | 1514 |
| (0.518,0.539] | 1380.910 | 1546.090 | 1356 | 1571 |
| (0.539,0.567] | 1309.782 | 1616.218 | 1254 | 1672 |
| (0.567,0.609] | 1210.871 | 1716.129 | 1155 | 1772 |
| (0.609,1] | 1001.566 | 1925.434 | 1060 | 1867 |

Table 5.8: Risultati del test di Hosmer-Lemeshow

Il p-value risultante dal test molto basso (0.0037) indica che l'ipotesi nulla deve essere scartata e quindi che il modello non è ben calibrato.

Si è provato a effettuare lo stesso test aumentando il numero di gruppi ottenendo p-value più alto, ma così facendo il test perderebbe di validità producendo dei range di probabilità insignificanti.

# Conclusioni

Riassumendo, per la realizzazione del progetto, nella fase di Data Understanding & Preparation, abbiamo assolutizzato il dataset originale facendo si che ogni record riportasse gli indicatori di una azienda in uno specifico anno. Dopo aver creato le variabili necessarie per effettuare le analisi richieste, abbiamo proseguito a rispondere ai task che ci sono stati assegnati.

Abbiamo prima valutato come la distribuzione dell'età e della dimensione cambia in maniera statisticamente significativa tra le aziende fallite e quelle attive in un determinato anno. L'analisi effettuata nel 2018, ha mostrato che tendenzialmente le aziende attive sono mediamente più giovani rispetto a quelle fallite. Invece le dimensioni delle aziende attive tendono ad essere superiori rispetto alle aziende non operative.

Obiettivo della seconda domanda, invece, è stato quello di valutare la distribuzione delle stesse variabili delle sole aziende fallite in un certo arco temporale. I test effettuati su campioni del 2010 e 2011 hanno riportato che le aziende fallite nel 2010 sono mediamente più giovani di quelle fallite nel 2011. Estendendo l'arco temporale, dal 2010 al 2014, il test di Friedman ci conferma la non similarità delle stesse. Per quanto riguarda la dimensione nei 5 anni, invece, la dimensione delle aziende fallite non è statisticamente differente.

La terza domanda richiede di analizzare in un determinato anno, la distribuzione dei fallimenti conoscendo l'età o la dimensione delle aziende. In entrambi i casi, abbiamo riscontrato un trend decrescente della probabilità di fallimento all'aumentare sia dell'età che della dimensione aziendale. Tutti questi casi sono stati analizzati anche al variare dei settori, delle forme societarie e delle regioni in cui le rispettive aziende operavano.

Per la risoluzione dell'ultima domanda, abbiamo allenato un regressore logistico non normalizzato, uno regolarizzato tramite l'utilizzo del metodo *Elasticnet* ed un modello di tipo *Random Forest*. Tutti i modelli sono stati allenati utilizzando un training set composto da dati relativi agli anni 2014-2015, poi testati sia su test set composti da dati relativi al 2014-2015 che da uno contenente quelli del 2013.
Dalla valutazione delle performance sul test set, è emerso che tutti i modelli classificano meglio record relativi al 2013. Relativamente alle performance ottenute dalle diverse tipologie di classificatori, abbiamo notato come Elasticnet riesca a migliorare (seppur di poco) le performance ottenute dalla regressione logistica non regolarizzata. In ogni caso, il modello Random Forest, è risultato migliore rispetto ad entrambi i regressori.