Statistical methods for data science
Project AY 2019/20

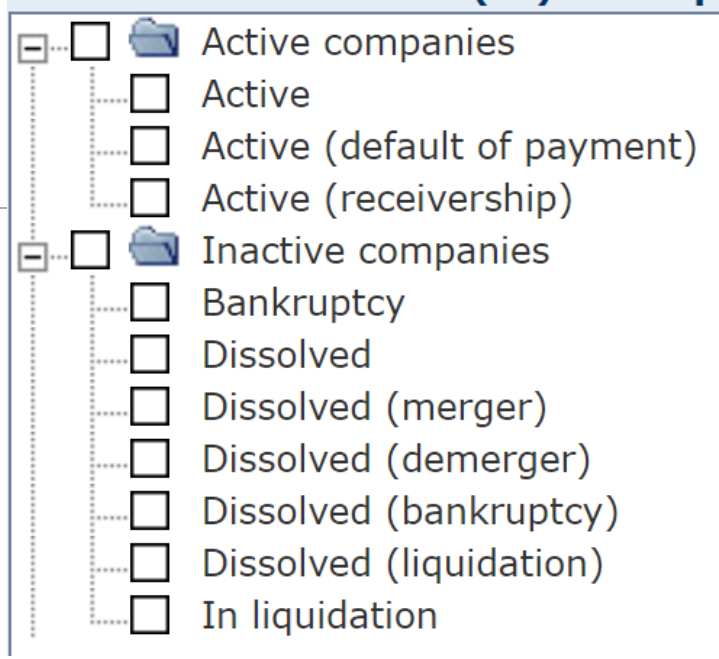
RISK OF BUSINESS FAILURE

Business going bad

- **Insolvency** is the state of being unable to pay the money owed, by a person or company, on time; those in a state of insolvency are said to be **insolvent**
- **Balance-sheet insolvency** is when a person or company does not have enough assets to pay all of their debts. The person or company might enter **bankruptcy,** but not necessarily.
  - If a loss is accepted by all parties, negotiation is often able to resolve the situation without bankruptcy.
- **Liquidation** is the process in accounting by which a company is brought to an end.
- An ended company is said **dissolved**.

Business Status

- **Active** – live and doing business
- **Active (default of payments)** – balance-sheet insolvency
- **Active (receivership)** – a trustee is legally appointed to act as the custodian of a company
- **Bankruptcy** – in the process of bankruptcy
- **In liquidation** – being closed (not for bankruptcy)
- **Dissolved** –closed



Bankruptcy/Failure prediction

Bankruptcy prediction consists of predicting bankruptcy and other financial distresses/losses.

- Will a firm go into a bankruptcy/liquidation/dissolved state?
- When it will happen?

Approaches:

- Parametric methods: curve fitting, statistical tests, regression, survival analysis, …
- Non-parametric (machine learning): decision trees, neural networks, ensembles, …

- **Preliminary setting**: fix a status subset as the notion of "failure"
  - Eg., Failure is Status == 'Bankruptcy'
  - E.g., Failure is Status != 'Active'

Question(s) (A)

Does the distribution of size/age change between failed and active companies?

- does it change for a specific company form (SPA, SRL, etc.)
- does it change for a specific industry sector? (see ATECO sectors)

i.e., there is a statistically significant difference

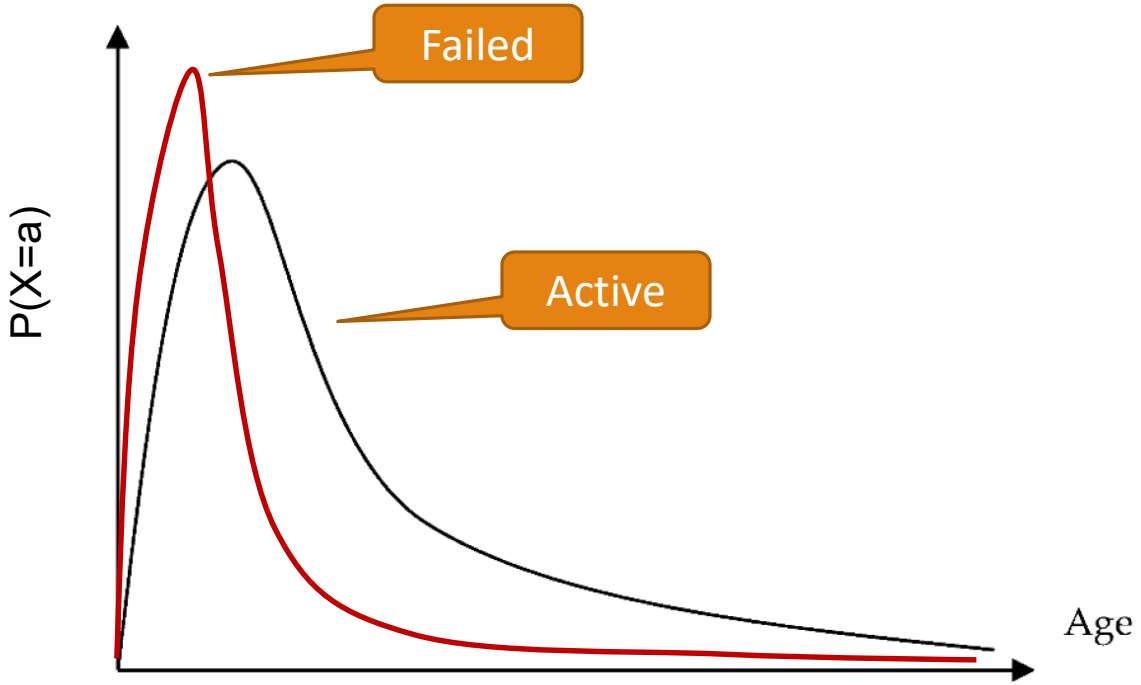


Figure 2. Age of company

Question(s) (A ctd)

Does the distribution of size/age change between failed and active companies?

- does it change for a specific location? (eg., Tuscany, Lombardy, etc.)
- does it change for a specific time? (eg., in 1/2/3 years from now)

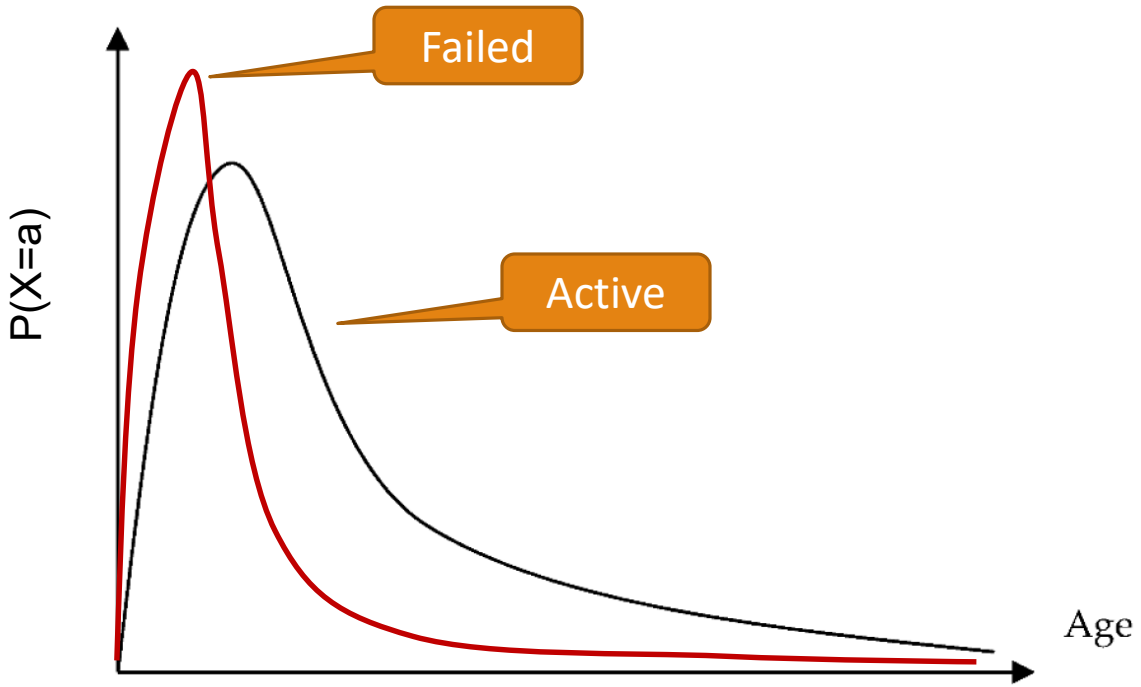


Figure 2. Age of company

Questions (B)

What is the distribution of failures wrt age/size of firms?

- does it change for a specific company form (SPA, SRL, etc.)
- does it change for a specific industry sector? (see ATECO sectors)
- does it change for a specific location? (eg., Tuscany, Lombardy, etc.)
- does it change for a specific time? (eg., in 1/2/3 years from now)

also, Probability of Default (PD) P(failure|X=a)

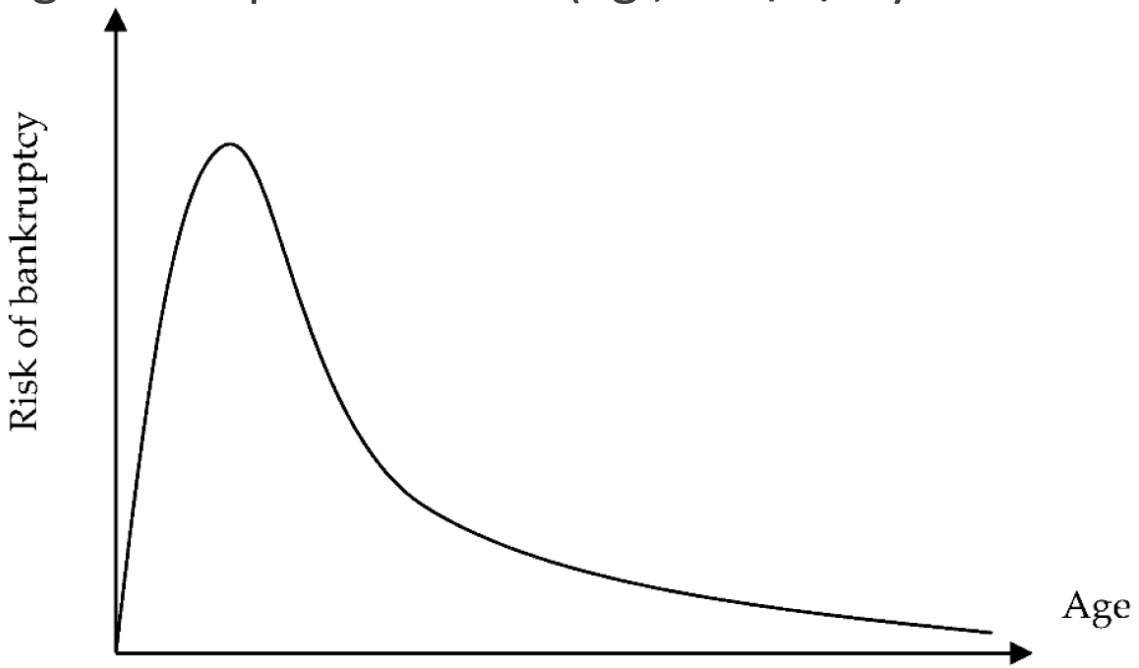


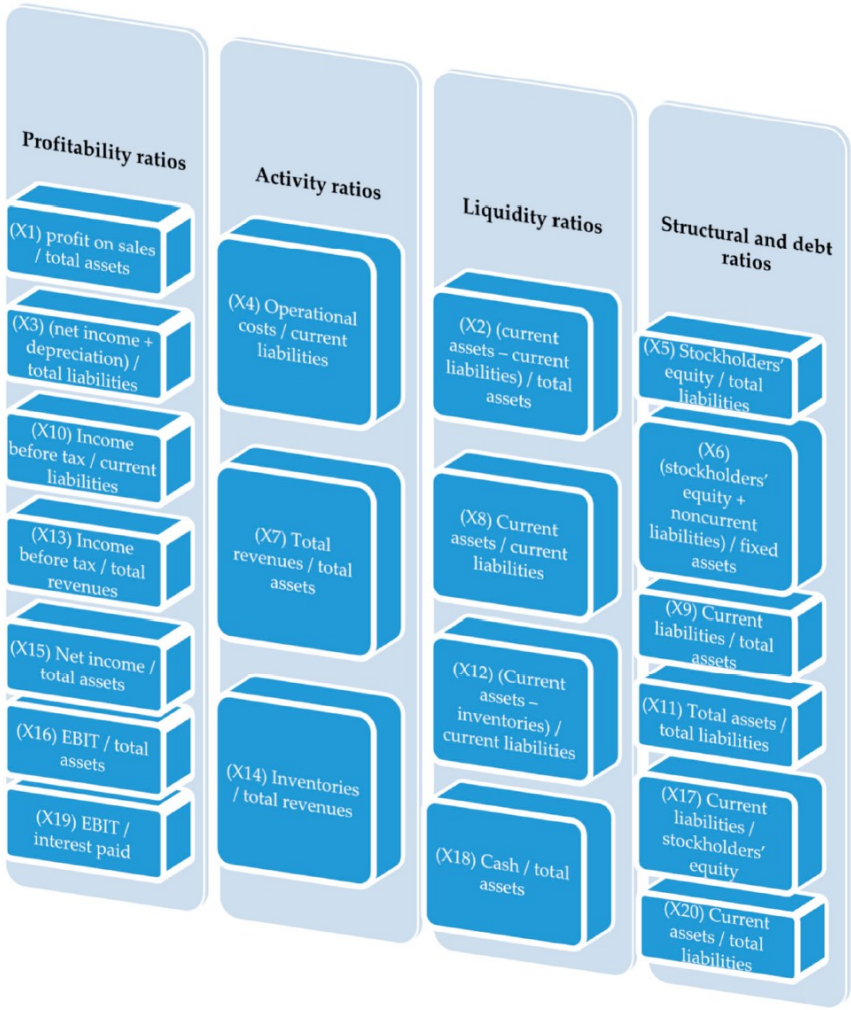
Figure 2. Age of company and the risk of bankruptcy.

Failure prediction

In addition to age, size, industry sector, and location, **financial indicators** that may correlate to failures have been widely studied in the literature, which motivates (credit/failure) scoring methods.

Table 1. Method used to compute selected financial indicators.

	Indicator	Computation method
CL	current liquidity	$\frac{\text{(current financial assets + current receivables)}}{\text{current liabilities}}$
TL	total liquidity	current assets/current liabilities
TATR	total assets turnover ratio	sales/total assets
ACP	average collection period	current receivables/sales × 360
CPP	creditors payment period	current liabilites/sales × 360
IR	indebtedness ratio	total debt/total assets
ER	equity ratio	equity/total assets
IC	interest coverage	EBIT/interests payable
ROE	return on equity	EAT/equity × 100
ROS	return on sales	EAT/sales × 100
ROA	return on assets	EBIT/assets × 100
CR	cost ratio	total assets/total revenues



Failure prediction

External data may also be part of the model (especially for multi-annual data):

- Market indexes (GDP, etc.)
- Financial indexes (ECB interest rates, etc.)
- Stock indexes (MIB, etc.)

Variable	Description
TURNOVER (Asset Turnover Ratio)	Ratio between net sales and total assets. The asset turnover ratio is an efficiency ratio that measures a firm's ability to generate sales from its assets.
VA_TA (Value Added to Total Assets)	Ratio between economic value added and total assets. Operating profitability ratio that measures the firm's ability to generate value from its assets.
EBITDA_MARGIN (EBITDA to Net Sales) - ***NL***	Operating profitability ratio that measures how much earnings the firm is generating before interest, taxes, depreciation, and amortization, as a percentage of revenue.
PFN/PN (Net Debt to Equity)	Measure of a firm's financial leverage, calculated by dividing its net liabilities by stockholders' equity.
EQ_TA (Equity to Total Assets)	Ratio between equity and total assets. Used to assess a company's financial leverage
PFN/EBITDA (Net Debt to EBITDA)	Debt sustainability ratio gives an indication as to how long a firm would need to operate at its current level to pay off all its financial debt.
IE_CASHFLOW (Interest Expenses to Cash Flow)	Ratio that indicates the enterprise's ability to pay interest from generated cash flow.
DSCR (Debt Service Coverage Ratio)	Ratio of debt sustainability that refers to the amount of cash flow available to pay interest expenses and annual principal payments on financial debt.
FIN_MISMATCH (financial mismatch)	Ratio of the mismatch (difference) between short-term liabilities and short-term assets and total assets. Negative value of the ration (short-term liabilities > short-term assets) indicates that the firm has enough short-term assets to meet its short-term liabilities.
CASH_ST_DEBT_S (Current Assets to Short Term Debt)	Liquidity ratio that measures a firm's ability to pay off short-term debt obligations with cash and cash equivalents.
CASH_TA (Cash to Total Assets)	Ratio between cash and liquid assets to total assets. It measures a firm's liquidity and how easily it can service debt and short-term liabilities if the need arises.
RECEIVABLES_TURNOVER (Receivable Turnover Ratio) - ***NL***	Efficiency ratio that measures how efficiently a firm is using its assets. It measures the number of times over a given period (usually a year) that a firm collects its average accounts receivable.
PAYABLES_TURNOVER (Payable Turnover Ratio) - ***NL***	Efficiency and liquidity ratio that measures how many times a firm pays its creditors over an accounting period.
LOG_ASSETS (Natural Logarithm of Total Assets)	Measures the size of the firm.
SALES_GWT (Net Sales Growth) - ***NL***	Measures a firm's growth in a specific year. It also measures the stability of a firm's performance.

Scoring methods

Parametric models (non-exclusive list):

- Linear Discriminant Analysis
  - Altman Z-score model

The model took the following formula:

$$Z\prime = 0.717 * x_1 + 0.847 * x_2 + 3.10 * x_3 + 0.420 * x_4 + 0.998 * x_5, \quad (5)$$

where x_1 is working capital/total assets, x_2 is retained earnings/total assets, x_3 is earnings before interest and taxes/total assets, x_4 is book value of equity/book value of total liabilities, x_5 is sales/total assets.

Zones of discrimination:
$Z' > 2.9 \rightarrow$ safe zone
$Z' \in < 1.23; 2.9 > \rightarrow$ grey zone
$Z' < 1.23 \rightarrow$ distress zone

- Logistic Regression

- Penalized (or Elastic Net Regularization) Logistic Regression

Questions (C)

Fit a parametric model, and use it for failure scoring:

- use one or more parametric models
  - you can rely on literature: just cite the source of the adopted approach
  - check whether the hypotheses of the model are satisfied (normality, multicollinearity, etc.)!
- split data into 70% training and 30% test
  - choose a quality measure to evaluate predictive performance
- explain in deep the results of the approach
  - particularly, what is the meaning of fitted parameters, confidence intervals, p-values, etc. in terms of failure prediction (eg., are young firms are more subject to failure than old ones all the rest being equal? etc.)

Machine Learning Models (Random Forests, Gradient Boosted Trees, …) can also be fit/compared **in addition** to at least one parametric model.

Resources

- **AIDA database** of (many) Italian companies
  - with historical data (last 10 years from closing)
  - https://www.sba.unipi.it/it/risorse/banche-dati/aida
  - **max 2** concurrent sessions!
- **Ateco 2007** classification of industry sectors
  - Italian version of the European NACE classification
  - https://www.istat.it/it/archivio/17888
  - Excel file + description notes
- Shared Google Drive directory
  - Require authorization
- Scientific paper indexes: Scholar, DBLP, arXiv, …

General rules

Teams work in **non-competitive** groups of up to 3 members

- Delivery: report (PDF) + code (.R)

Data collection from AIDA website can/should be shared among teams in the Google Drive directory

- Teams+Professor should agree on a set of attributes to collect
- and distribute downloads among them
  - E.g., by assigning distinct Italian regions to each group

Timeline and deadlines

- Group formation: by 24 March 2020
- Data collection: by end of April 2020
  - Status of project will be monitored in classes during the semester
- Delivery of report: by end of July 2020
- Oral discussion: by 12 September 2020
  - Oral includes discussion of the project and open questions on the topics of the course