Assignment M2 DS2E – 2023/24

Context: Detecting high-growth firms

The data that you are going to use for this project comes from a random extraction of the Survey on Business Strategies (Encuesta Sobre Estrategias Empresariales – ESEE). This annual survey gathers extensive information on around 2,000 manufacturing companies operating in Spain and employing at least ten workers. The dataset at your disposal covers the period 2000–2012.

1990-2012

With this type of information, a data analyst can finally start looking for (potentially interesting) correlations and patterns regarding what determines high-growth firms (HGF). HGF can be defined in two different ways: (i) based on their sales growth rate [continuous variable]; (ii) companies with extreme growth performance, i.e. belonging to the top 10% of the annual sales growth distribution [binary variable]. Faire 2 modèles

Fixer un seuil

Your task is predict which company is going to be a HGF in the last of year of the sample [your test set]. This is very important for managers and policy-maker, for example if they wish to identify promising companies for M&A and/or target financial support.

It means we can drop
companies which have NA in
the last years

Here is the description of the variables at our disposal:

● *id*: Company's ID [anonymized]
● *year*: Year
industry: Company's main sector of activity [anonymized]
yestab: Year in which the company was established
● *pertot*: Number of employees
● *enggrad*: Share of engineers and graduates
● *sales*: Total sales
● *va*: Value added
● *gom*: Gross Opearating Margins
● *rdint*: Internal (in-house) R&D expenditures
● *rdext*: External R&D expenditures
● *ipnc*: Product innovation - New components [dummy]
● *ipnf*: Product innovation - New functions [dummy]
● *ipnm*: Product innovation - New materials [dummy]
● *ipr*: Process innovation [dummy]
● *patent*: Number of granted patents

sales_growth_rate

growth_top_10

Here some suggestions. Pre-process of data: NAs, outliers, etc.. Create new features from the available information, such as R&D intensity, age, industry dummy, etc.. Try to understand the data through some visualisations before modelling.

You are expected to provide your Python code/notebook, clean and well commented, and present the results of your analysis with the support of max. 5 slides in 5 minutes.