A Machine Learning approach to classification of companies

Fabio Morea

2022-02-04

Business case

RCM is the regional association of composite material processing companies, that counts over 1000 associates (companies that have their head office or a local unit in Friuli Venezia Giulia and are interested in some way in manufacturing with composite materials). RCM is monitoring the performance of its members: each company with a performance level (Top/Mid/Low) that is used for further activities (e.g. Top performers are featured in the newsletter, Mid and Low performers receive different proposals…). The classification algorithm is based on financial ratings issued by Modefinance (purchased at 5,00€ per company). Rating is expressed as a value 1 to 10 and The performance level is now based on ModeFinance rating:

- Top: rating 7 to 10
- Mid: rating 5-6
- Low: rating 1 to 4

Objective: RCM wants to develop a new method, leading to the same classification, replacing financial ratings with balance sheet data (which can be purchased at a significantly lower cost, € 0.75 per company). The expected result is a model (learning and prediction modules) and a detailed report describilng the model performance (error rate, robustness to unbalanced data, . . .)

Constraints: training and classification will be performed on a laptop, twice a year. No specific constraints on time or computation effort (even if it takes hours, it's ok).

Workflow - insert picture here - Classification tree

Understanding the data

TODO insert figure here

Data is available from 2 sourecs: * cmp.csv * bsd.csv * rating.csv

and sould be pre-processed to obtain two vectors: X and y.

About this notebook

The notebook is divided in X sections: an introduction, a section for each dataset and a final section on potential future development.

1. Companies and local units
2. NACE activity sectors
3. Financial indicators and rating
4. Employees: stock and flow

5. Conclusions and further development

The notebook has been written using *R-Studio* and rendered with *boowdown* (https://bookdown.org/) package.

Data data manipulation is based on *tidyverse* [https://www.tidyverse.org/], a data science library that includes *magrittr* (pipe operator %>%), *dplyr* (select, summarize…), *tibble* (a tidier version of the data.frame) and *ggplot2* (visualizations). A useful guide to tidyverse is available online at the following address: [https://r4ds.had.co.nz/]

Data management plan

Raw data: The original data has been pre-pocessed by Area Science Park to fulfill the following requirements:

- encoded in UTF-8 cleaned from non-printable characters
- table columns are attributes (features, independent variables), renamed to be human- and machine-readable
- table rows are observations If you have multiple tables, they should include a column in the table that allows them to be linked
- splitted into several tables, created unique identifiers to connect the tables
- saved each table to separate .csv file with a hunam-readable name

No attributes were removed or summarized during pre-processing. Raw data is available in local folder *data/raw*.

Tidy data: This notebook explores all the attributes available in the raw data, and by merging, subsetting and transforming, produces a smaller, cleaner data set ready for further analysis. Tidy data is saved in local folder *data/tidy*; examples of groups of companies used for data exolporation are saved in subfolder *data/tidy/groups*.

This notebook describes the process used to transform raw data into tidy data, the contents of individual files, the meaning of each attribute and examples of visualization. A basic feature selection is also introduced, prunin attributes that are not relevant for future reseach.

Updates: raw and tidy data may be updated periodically, since Area Science Park updates the raw data set twice a year; anyway the current version in based on June 2021 version and does not provide automatic updating scripts.