A Machine Learning approach to classification of companies

Fabio Morea

2022-02-05

Scope and objectives

This collection of notebooks presents some excercises for the course “introduction to machine learning”.

Background information:

Research, innovation and highly skilled people are considered to be important factors in economic and social development. Economic support policies often include funds to support research (for example with the creation of public research infrastructures), companies (for example with tenders to co-finance innovative projects) and the training of people with the necessary skills.

Area Science Park is a national research institution that manages a science and technology park located in Trieste (Italy) and is engaged in several projects aiming to support innovation at the regional level.

Innovation Intelligence FVG is a project, managed by Area Science Park and supported by Regione Friuli Venenzia Giulia, which aims to monitor the performance of companies in terms of economic, employment and innovation results. For more info refer to: (www.innovationintelligence.it/).

Target groups include the “tenants” (60 companies that have their premises or research laboratories in the science and technology park), or other groups if companies identified by some project (for example the “regional cluster” of small and mediom entermrises working in the field of plastic manufacturing).

The core result of Innovation Intelligence FVG is a dataset containing information from several data sources such as the chamber of commerce, the Regional Labor Market Observatory, a rating agency, as well as regional databases on innovation projects. Most of the data is open sourced; company information is available trough the chambers of commerce at a cost of 0,30€ per company, while “financial ratings” are the most expansive part, at an average cost of over 1,00€ per company.

Objective

The objective of this project is to introduce more advanced analysis techniques based on the data available, and to test the potential replacement “financial ratings” with freeo or less expansive data sorucs.

Case study

The new analysis techniques should be focused on the following case study: RPM the association of plastic 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 plastic materials). RPM is monitoring the performance of its members, assigning each company 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 is currently based on financial ratings (a numeric variable in range 1 to 10):

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

RCM wants to apply a new method, possibily based on machine learning, that result in a classification leading to the same classification in (Top/Mid/Low), based on "balance sheet data" (which can be purchased at a significantly lower cost, € 0.30 per company).

The **target audience** is a small group of economic analysts, that have a robust domain knowledge but limited experience in data science.

The **expected results** ares a model (learning and prediction modules) and a detailed report describing 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). The number of companies involved is of the order of 1000.

Data management plan

The original data available from Innofation Intelligence fulfills 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.

Pre-processing is described in a separate notebook, providing furhter details on all the attributes available in the raw data, and the transformations used to produce a smaller, cleaner data set ready for further analysis. Tidy data is saved in local folder *data/tidy*.

Original and tidy data are updated on a monthly basis; the current version in based on June 2021 version and does not provide automatic updating scripts.

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).