Prediction methods - mid-term evaluation

Pierre Michel

MASTER in Economics - Track EBDS - 2nd Year (2023)

Using the XGBoost algorithm

The aim of this project is to describe how the XGBoost[1] (extreme gradient boosting) algorithm works and to test its implementation in Python (or R, as you prefer) based on an example dataset.

This project will require you to search for information on your own, and ask yourself the following 3 questions:

- For what type of machine learning task is this method useful (supervised versus unsupervised, classification or regression)?
- How does the algorithm work?
- Are there any applications of this method to economic or social science data?

At the deadline, you will be asked to submit **a written document of max 10 pages** presenting the following sections:

- *Introduction*: in this section you should briefly present the algorithm, give references to the literature on XGBoost and its applications in economics and the social sciences, and present an idea of how it might be applied to data (the choice of data is up to you).
- *Materials and methods*: in this section, you will describe the data used for your application, explain how XGBoost works, propose an application on your data, describe the problem and discuss model selection and validation.
- *Results*: this section should present a statistical description of the data used, as well as the results of your algorithm in terms of prediction performance (the performance metrics should be chosen appropriately according to the data and your objective).
- *Conlusion*: this should include a discussion of the advantages and disadvantages of XGBoost, and how it compares with other decision-tree based methods seen in class.

A notebook should also be supplied with the written document, containing the code for reproducing your work. The evaluation will take into account various aspects, including the clarity of the document and the code, as well as the care taken with model fitting and validation.

[1] Chen, Tianqi, et Carlos Guestrin. 2016. «XGBoost: A Scalable Tree Boosting System». In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-94. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785.