Econometrics & Machine Learning

Amal Elfassihi, Nicolas Khadivi, Bastien van Delft

December 9, 2016

1 Background

This project aims to investigate how Machine Learning algorithms (e.g. decision trees, random forests, boosting) compare to more traditional econometric tools such as linear or logistic regressions, using a real-world dataset [1] to study a binary variable with these various algorithms.

Our objective is twofold:

1. compare prediction quality between the various algorithms
2. compare interpretability of the models and recover marginal effects if possible

This project is supervised by Dr. Romain Aeberhardt and Thomas Larrieu.

2 Current progress

2.1 Research

On the advice of our supervisor, we have researched the theoretical background of the project. More specifically, we have looked into (and learned) basic Econometrics methods that we will need to apply throughout the project, using [2], and in conjunction with the Econometrics I class taught by F. Kramarz. We have also started learning some Machine Learning methods, understanding the basics using online material [3], and then studying further with the help of a Machine Learning textbook [4].

2.2 Setup

As this project will involve a fair amount of programming, we have also ensured that we have an adequate experimental setup that will allow us to implement the algorithms and experiment in a relatively flexible way, yet also let us record and analyse our results with relative ease.

Our experimental setup revolves around various tools that we have put in place:

- the Python and R programming languages, along with the scikit-learn machine learning package [5][6][7]
- the Git source control manager [9], which enables us to keep track of different versions of our code (and thus project), while also letting us easily rollback changes, or run multiple versions of the code (e.g. with different parameters)
- Jupyter [8], an interactive computing platform, which enables us to run code directly in the browser, and mix text with code snippets to document our work properly

We have also downloaded the dataset we intend to work with, *Enquête Emploi en Continu* [1], performed some basic manipulations to make it easier to work with (conversions), and started investigating the kinds of data we have access to, as well as running some very basic regressions (more for practice than for finding useful results).

Our work can be found at all times on the project's GitHub repository and in the Jupyter Notebook.

3 Future

Over the next weeks (starting after the exam period) and with the agreement of our supervisor, we intend to advance the project by performing the following steps:

1. explore the database further to understand what kind of data we have, and what parameters we would like to include in our models; this entails investigating the various parameters, through summary statistics and graphical representations
2. run various classical multiple linear regressions with adequately chosen parameters, to understand the effects of each of them on our dependent variable (probability of `employed`/`not employed`)
3. run logit and probit regressions, and understand the differences with a regular MLR regression
4. move on to machine learning algorithms (decision trees, then random forest and boosting), and investigate ways in which we can extract marginal effects

References

[1] INSEE, *Enquête Emploi en Continu*. 2015

[2] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, CENGAGE Learning, 2015

[3] AnalyticsVidhya.com, *Essentials of Machine Learning Algorithms (with Python and R Codes)*. 2015

[4] G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2013

[5] The Python Programming Language.

[6] The R Project.

[7] scikit-learn: machine learning in Python.

[8] Jupyter Notebook.

[9] Git SCM.