Computational Statistics:
Data Science for Social Science

Lena Janys
Assistant Professor Institute for Finance and Statistics
ljanys@uni-bonn.de

April 11, 2022

Syllabus:

Zoom Link: https://uni-bonn.zoom.us/j/68082760734?pwd=Z1dKZHJSVDM2OThVNXVNblNoZFBTUT09

The goal of this course is to

1. Introduce basic methods from data science and machine learning.
2. How to apply these to problems in empirical social science.

This is a very applied course, where we want to emphasize how to *implement* a particular method and how to *interpret* the results.

To this end, the course will be divided into three blocks:

1. Introduction to R:
2. Basic Concepts of Data Science:
   (a) Classification
   (b) Validation
   (c) Structural Estimation vs. Prediction
   (d) Resampling Methods
3. Methods in Data Science
   (a) Regularization: Principal Components, ridge regression, lasso
   (b) Regression Trees, random forests, bagging, boosting stacked methods/ensemble methods
   (c) Causal forests
   (d) Deep Learning/Neural Networks

I will interweave applications from social science, as well as common pitfalls when implementing these methods and particular problems that occur in social science applications.

Literature and languages:

1. Programming: we will use R for all implementations. Please download any distribution of R for your machine `https://cran.r-project.org` and you may also download any other editor than the build-in one, I prefer `https://rstudio.com/products/rstudio/`.

2. The main reference will be James et al. (2021), which is available for free as a PDF. I will subsequently add more literature as we discuss specific applications. Further reading is Friedman et al. (2001)

Format:

- The lectures will be held live at the times specified in BASIS on Tuesday and Wednesday.
- All materials will be available online on the GitHub repo `https://github.com/LJanys/CompStat`, (hopefully very soon after the lectures). I will also upload all other materials on this repo.
- If you cannot be present during the lecture, please contact me and we will find an arrangement.
- Tuesdays are reserved for implementation and presentation of problem sets. Once we finish the R introduction/ revision of some econometric concepts, I would like for everyone to present their results of the problem set by sharing their screen and discussing the results in class, so please expect to do that at some point and to participate actively in class, even if you are not presenting. This means (1) be prepared (2) think of one or two questions you would like to ask the presenters and/or that you had while working on the problem set. Be prepared to be called to ask those questions at the end of the presentation.

Grade: Instead of an exam there will be a project/Hausarbeit, for which you will need to pick a topic in a couple of weeks, once you have decided that you want to stay in the course. You need to register for the course with the Prüfungsamt between **13.04.– 20.04. at the latest**. You will also be asked to present (parts) of a problem set, possibly in groups. Once everyone has registered for the course, you will be assigned a problem set to present (possibly in groups). 10% of your grade will consist of that presentation. The problem sets will involve implementation of different methods in R and interpretation of the results.

Project: The deadline for the individual project will be August 23th, as the deadline has to be least 4 and at most 6 weeks after the topics were approved/assigned (this is stipulated by the examination office). This would mean that the topics will be finally chosen in the beginning of July, where we will do a "mini-workshop", where everyone introduces their topic and we can discuss some issues that might arise. The project will involve a simulation study with a realistic empirical set-up, and (ideally) an empirical application of one of the methods discussed in class in a social science setting.

References

Friedman, J., T. Hastie, and R. Tibshirani (2001): *The elements of statistical learning*, vol. 1, Springer series in statistics New York.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021): *Statistical learning*, Springer.