Prof. Dr. Daniel Schnitzlein **Institute of Labour Economics**

M.Sc. Vangjel Bita **Institut für Arbeitsökonomik**

Seminar Data Analytics, Winter Semester 2023/24

Seminar Description

In this seminar, we invite students to engage in practical exercises, gain hands-on experience with Python, and apply their machine-learning skills within the field of economics. Students will form teams and be assigned projects, which include a specific topic and dataset to work on throughout the semester. Additionally, python code and helpful instructions will be provided to start with. (Programming experience is advantageous but not necessary,). At the end of the semester, students will present their work in front of their peers. Our research projects will focus on leveraging the predictive advantages offered by machine learning in empirical economic research designs. Specifically, we will utilize machine learning techniques to:

1. Estimate the gender wage gap in weekly earnings with partial linear regression where ML predictions are used to find the optimal controls.
2. Explore heterogeneity in treatment effects arising from experimental data from cash transfers and child health in Indonesia
3. Make accurate predictions of the risk of repeating a crime of individuals that have been already convicted once.
4. Predict occupations' risk of automation based on text data of task descriptions of each occupation.

Learning Goals

Programming in Python. Hands-on experience with data manipulations, visualizing and plotting results, and training machine learning algorithms.

- Getting familiar with working and interpreting ML models.
- Using ML methods for best predictions.
- Using ML for causal analysis, choosing best controls, and estimating treatment effect heterogeneity.
- Working with text as data.
-

Aim

The idea is to turn students into knowledgeable users of these methods with awareness of their potential pitfalls when transferring methods to economic applications. Such pitfalls can be data quality and bias, over-fitting, causal inference and endogeneity, interpretability, and transparency behind the algorithms, model complexity, and ethical considerations. The course aims to familiarize students with ML algorithms, work on applications in economics based on the latest research and support them to work independently on a project in Python.

Course Structure

The seminar will consist of three meetings and the following deadlines.

- Seminar introduction. Team formation and topic assignment. Date: 13.10.2023
- Consultation meeting for feedback and questions. Date: 08.11.2023
- Presentation round. Date: 30.11.2023.

Students will be provided with one of four datasets and useful code to start with. They will be provided via a link to the course's StudIP repository. To each team, one of the following dataset and ML method/model will be assigned.

1.Data CPS Weekly Earnings → estimate the gender gap (Dim:302.332x98)

2.Data Margaret Triyana Heterogenous Treatment effects Indonesia AEA web (Dim:22.771x121)

3.Data Recidivism predictions of Risk of Recidivism (Dim:7.214x53)

4.Frey, Osborne the Future of employment (2013) → Text data for prediction of risk of automation.

Books and Literature

Relevant literature per data Group:

1.Data Group 1: Double/debiased machine learning for treatment and structural parameters.

2. Data Group 2: Do Health Care Providers Respond to Demand-Side Incentives? Evidence from Indonesia

3. Data Group 3: Compas risk scales: demonstrating accuracy equity and predictive parity, Human Decisions and Machine Predictions

4. Data Group 4: The future of employment: How susceptible are jobs to computerization?

Behind the Headline Number: Why not Rely on Frey and Osborne's Predictions of Potential Job Loss from Automation

Further Books.

1. Friedman, Hastie, and Tibshirani: Elements of Statistical Learning

2. Goodfellow, Bengio, and Courville: Deep Learning

3. Efron and Hastie: Computer Age Statistical Inference

Methods

1. Random Forests
2. SVM
3. Neural Networks
4. Text analysis

Examination

1.Submission of Code Script: Code that reproduces graphs tables and the results of your analysis in R or Python.

2.Term Paper: 8 Pages max of Data visualization, descriptive, method, choice of variables, training and tuning choices, and interpretation of the results. In the term paper, the benchmark regression method should be analyzed to be compared with ML and showcase its advantages or disadvantages.

3.Presentation: 15 Minutes explaining the ML Method theory, Variable Choice Reasons, Training tuning steps explained, and results.

4.Grade Determination:

– Seminar Thesis: 45%

– Oral Presentation: 45%

– Active Participation in the Seminar:10%

Important Dates and Deadlines

1.Registration date: 12.10.2023

2.Introductory meeting for groups, and topic distribution. Date 13.10.2023

3. Feedback meeting 2 weeks before submission for final questions and guidance. Date 08.11.2023.

4. Seminar Presentations: 30.11.2023

5. Term Paper submissions: 23.11.2023