Swiss Leading House VPET-ECON

A Research Center on the Economics of Education, Firm Behavior, and Training Policies

University of Zurich[UZH]
Co-Director
Prof. Dr. Dr. h.c. Uschi Backes-Gellner

UNIVERSITÄT BERN
Co-Director
Prof. Dr. Stefan C. Wolter

Fall Term 2023

Leading House Ph.D. Course
"Machine Learning with Python for Education and Personnel Economists"
- Syllabus -

INSTRUCTOR	**Michael E. Rose**, PhD, Post Doctoral Researcher at Max Planck Institute for Innovation Competition, Munich, Germany
Workshop dates	October 9 – October 13, 2023
Location	University of Zurich, Room tba
Preliminary Schedule	The lectures take place in the form of an intensive 5-day course. Monday, October 9: 12:00-18:00 Tuesday, October 10: 08:00-18:00 Wednesday, October 11: 08:00-18:00 Thursday, October 12: 08:00-18:00 Friday, October 13: 08:00-12:00
Module Number, ECTS	DOEC0918; 3 ECTS
Course Webpage	http://www.business.uzh.ch/professorships/emap/teaching.html

Course Description

The course allows doctoral students in Economics of Education and/or Personnel Economics to start their own empirical Python projects and introduces various Machine Learning models. It also teaches some highly relevant python packages, version control, and project management. Most units are highly practical and complemented with a training session.

Prof. Dr. Dr. h.c. Uschi Backes-Gellner
University of Zurich, Department of Business Administration, Plattenstr. 14, CH-8032 Zürich
Tel. +41 44 634 42 81, backes-gellner@business.uzh.ch

Prof. Dr. Stefan C. Wolter
University of Bern, Department of Economics, Centre for Research in Economics of Education,
Schanzeneckstr. 1, Postfach 8573, CH-3001 Bern
Tel. +41 31 631 36 56, stefan.wolter@vwi.unibe.ch

Outline

Unit 0: General Introduction
- Getting to know each other
- Learning outcomes, course outline, explanation of examination
- Some python background

Unit 1: Python Project Management
- The PyCharm IDE
- Introduction to pandas, matplotlib and seaborn
- Version control with git
- Collaborating with GitHub/GitLab
- Debugging

Unit 2: Introduction to Machine Learning
- Why Economists & Econometricians should know about ML methods
- Definitions & Algorithms
- Possible (future) applications
- [Feature engineering]

Unit 3: Unsupervised Machine Learning
- Examples
- [Principal Component Analysis]
- Scaling
- Agglomerative clustering, Hierarchical clustering, DBSCAN
- Cluster evaluation metrics

Unit 4: Supervised Machine Learning
- Examples
- Workflow I: Split, train, evaluate
- Distances and evaluation metrics
- Linear models and regularisation (Ridge, LASSO)
- Neural networks
- Workflow II: Cross-validation, pipelines, grid search
- Machine Learning for Econometricians

Unit 5: Traditional Natural Language Processing
- Examples
- text cleaning and encoding problems, stemming
- Text vectors
- Sentiment analysis with TextBlob; Readability analysis

University of Zurich[UZH]
Co-Director
Prof. Dr. Dr. h.c. Uschi Backes-Gellner

UNIVERSITÄT BERN
Co-Director
Prof. Dr. Stefan C. Wolter

Unit 7: Advanced Natural Language Processing

- Embeddings
- BERT family
- Excourse on ChatGPT
- Excourse on regular expression
- Excourse on Latent Dirichlet Analysis

Unit 8: Outlook

- [dedupe]
- Python Style Guide PEP8
- My own workflow
- Packages for advanced ML
- [Info re Examination]

Required preparation:

Basic knowledge in Python is a requirement to participate in the course.

- If you never worked with Python before (non-experienced Python-User): Finish the self-paced, interactive online courses for Python (about 2 hours) at https://www.udemy.com/course/learn-python-3-from-scratch-python-for-absolute-beginners/ and forward a screenshot of the last page showing your name (alternatively, the certificate you receive in the end) to Michael.Rose@ip.mpg.de
- If you worked with Python before, please briefly state your previous experience and knowledge in Python to Michael.Rose@ip.mpg.de

In both cases, the deadline for the confirmation of your Python credential is **Saturday, October 7, 2023**.

Non-experienced Python-Users are required to get up to speed by a Zoom-session from the lecturer on **Sunday, October 8, 2023, at 7 pm**.

Required readings:

- Shapiro, J. and M. Gentzkow: "Code and Data for the Social Sciences: A Practitioners Guide"
- Athey, S. and G. Imbens (ARE 2019): "Machine Learning Methods That Economists Should Know About"
- Gentzkow, M., B. Kelly and M. Taddy (JEL 2019): "Text as Data"

Required software:

- (Windows users:) Python 3.7 or higher - during installation, make sure both python and pip are part of the PATH (while installing Python, check "Add Python 3.X to PATH", then click "Customise Installation" and check "Add pip to PATH"). Note: the use of anaconda is discouraged (see Notes below).
- (All users): install a modern text editor of your choice, e.g., Sublime Text 4 or Notepad++ (only Windows); Notepad will not work
- (All users:) PyCharm Community Edition (default settings are recommended)
- (All users:) git (select the editor from step 2 as default editor) - optionally, add any of the GUI clients

Teacher

Michael E. Rose, PhD; Post-Doctoral Researcher at the Max Planck Institute for Innovation and Competition, Munich, Germany (Department "Innovation and Entrepreneurship", headed by Dietmar Harhoff)

- Teaching experience:
    - This course (Python and Machine Learning, for PhD/Dr students) at TUM School of Management, Kiel Institute for the World Economics, U Zurich/ETH Zurich, LMU Munich, ifo institute, and Georgia Tech's Scheller College of Business
    - Computational Mathematics (Matlab, SQL, VBA, Excel) for Master's students: at U Cape Town
    - Time Series Econometrics (Python, for Master's students) at U Cape Town
- Daily usage of Python, ML in own research
- Lead development of two python packages (pybliometrics and sosia)
- Co-development of Logic Mill (https://arxiv.org/abs/2301.00200)

- Co-PI in Academic Research Project "Tracing the flow of knowledge from science to technology using deep learning" 2021 of the European Patent Office

Related literature:

- Downey, Allen B.: "How to think like a Computer Scientist", Chapters 1, 2, 3, 5, 8, 10, 11, 12
- Shapiro, J. and M. Gentzkow: "Code and Data for the Social Sciences: A Practitioners Guide"
- Gentzkow, M., B. Kelly and M. Taddy (JEL 2019): "Text as Data"
- Mueller, A. and Sarah Guido: "Introduction to Machine Learning with Python"
- Athey, S. and G. Imbens (ARE 2019: "Machine Learning Methods That Economists Should Know About"
- Mullainathan and Spiess (JEP 2017): "Machine Learning: An Applied Econometric Approach", 31 (2).
- Athey, S. (Science 2019): "Beyond Prediction: Using Big Data for Policy Problems".

Notes:

- We are going to use **plain** Python 3, while I recommend not to use anaconda. On the first day I'll explain why. Anaconda users might not be able to replicate everything.
- The Integrated Developer Interface (IDE) I am going to use and show is **PyCharm** Community Edition. spyder is also fine, but as it offers less functionality, I cannot offer support for spyder.
- Recommended text editors: **Sublime Text 2, Notepad++**. Do not use Windows' native editor Notepad.
- You are going to be assessed based on code you submit via a private shared GitHub repository. Signing in on **GitHub** is necessary for this (and free, of course), so you might want to do this beforehand. If you register for the first time, pick a proper username, as you might want to make use of the account in your professional career.

Target audience

The course is particularly designed for doctoral students in the course programme on economics of education of the Swiss Leading House. Doctoral students in economics or business economics with a strong interest in economics of education or personnel economics and machine learning are welcome as well. The seminar will take place en bloc in order to enable external Ph.D. students to attend.

Credit Requirements and Grading

1. Full course attendance. Students are expected to come prepared to class and exercises. It will facilitate discussion and improve overall learning.
2. Each course unit is complemented with a training session that takes place in the afternoons. Each student has to finish all the exercises presented during the training sessions at the end of the week (Sunday, October 29, 20:00 CET). The work is to be done in groups of two (or three if there is an odd number of participants).

Application

The number of participants is limited. Course registration until **September 4, 2023,** is mandatory. Please send your application (including a one-page CV) to chiara.zisler@business.uzh.ch.

WWF Course Policies

According to WWF study regulations, all exam dates are final as published in the VVZ and syllabus. This means that the final exam date is not negotiable. It will not be possible to take any exams on different dates.

Academic dishonesty in any form will not be tolerated. Anyone caught cheating or engaging in unethical behavior will be reported to the Dean's office according to the guidelines on academic dishonesty set forth by the University of Zurich.

The information in this syllabus supports the official information in the electronic university registration tool (VVZ – Vorlesungsverzeichnis). In cases of doubt, the official information at the VVZ is decisive.

For UZH students: Don't forget to officially register using the registration tool of the University of Zurich.

Any work presented by students only uses aids that are declared. Students accept full liability for the scientific integrity irrespective of potential use of generative AI tools such as ChatGPT. All instances of direct quotes or paraphrasing from published or unpublished sources are properly attributed. All work, in its current form or any similar version, has not been previously submitted, in whole or in part, as part of any other examination.