HUMBOLDT-UNIVERSITÄT ZU BERLIN

Daniel Jacob

Humboldt-Universität zu Berlin
School of Business and Economics
Dorotheenstr. 1
10178 Berlin
☎ *+49 30 2093 99468*
✉ *daniel.jacob@hu-berlin.de*
hu-berlin.de/jacob

Doctoral Studies

2018– Present **Humboldt-Universität zu Berlin**, *PhD*, Information Systems and Statistics, Expected completion July 2021. DISSERTATION: "Essays in Econometrics: Causal Inference and Machine Learning" .

DISSERTATION COMMITTEE AND REFERENCES:

Prof. Dr. Stefan Lessmann
Chair of Information Systems
Humboldt-Universität zu Berlin
+49 30 2093-99542
stefan.lessmann@hu-berlin.de

Prof. Dr. Wolfgang Karl Härdle
LvB Professor of Statistics
Humboldt-Universität zu Berlin
+49 30 2093-99592
haerdle@wiwi.hu-berlin.de

Prof. Dr. Qingliang Fan
Department of Economics
Chinese University of Hong Kong
(852) 3943-8001
michaelqfan@cuhk.edu.hk

Prior Education

2018 **Master of Science in Economics**, *Humboldt-Universität zu Berlin.*

2015 **Bachelor of Science in Economics**, *Humboldt-Universität zu Berlin.*

Research and Teaching Interests

Causal Inference, Machine Learning

Publication

"Affordable Uplift: Supervised Randomization in Controlled Experiments" with Johannes Haupt, Robin Gubela and Stefan Lessmann, *International Proceedings in Information Systems (2019)*

Working Papers

"Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects" (IRTG Discussion Paper)

"Group Average Treatment Effects for Observational Studies" (IRTG Discussion Paper)

"Heterogeneous Treatment Effects through Tenure on Job Satisfaction: A Machine Learning Evaluation"(joint work with Qingliang Fan and Sue Ge)

Invited Conferences

2020
- American Causal Inference Conference (ACIC), Austin, Texas, U.S. (postponed)
- Causal Machine Learning Workshop, Einstein Congress Centre, St. Gallen (Switzerland)

2019
- Workshop in Microeconometrics, WISE, Xiamen University, Xiamen (China)
- AI and Data Science Workshop, National Cheng Kung University, Tainan (Taiwan)
- Workshop in Empirical Economics, Universität Potsdam, Berlin (Germany)
- Stat of ML Conference, Charles University, Prague (Czech Republic)

Awards, Fellowships

- CENTRAL Kollegs Workshop, Fellowship 2019
- Best presentation award at Statistics of Machine Learning Conference, Berlin 2019
- Deutsche Forschungsgemeinschaft via the IRTG 1792 "High Dimensional Non Stationary Time Series" Scholarship, 2018-2020

Programming

R (expert), Python (intermediate), LaTeX(expert)

Experience

11.2020–04.2021 **Teaching Instructor**, *IRTG 1792*, Humboldt Universität zu Berlin.
Semi- and Nonparametric Modelling (graduate and PhD students)

10.2019–02.2020 **Visiting Researcher**, *WISE*, Xiamen University, China.

2019–2021 **Teaching Instructor**, *Chair of Information Systems*, Humboldt Universität zu Berlin.
Seminar Applied Predictive Analytics: Causal Inference and Machine Learning (graduate students)

10.2018–04.2019 **Teaching Instructor**, *Ladislaus von Bortkiewicz Chair of Statistics*, Humboldt Universität zu Berlin.
Seminar Digital Economy and Decision Analytics (graduate and PhD students)

03.2017–09.2018 **Teaching Assistant**, *Ladislaus von Bortkiewicz Chair of Statistics*, Humboldt Universität zu Berlin.
Statistics II (winter term 2017/18 and 2018/19); Statistics I (summer term 2017 and 2018); Shiny Apps programming in R

01.2016– 04.2016 **Research Assistant**, *Oxford University Press*, London.

04.2015– 04.2017 **Student Assistant**, *Institute for Economic Theory II*, Humboldt Universität zu Berlin.

Research Papers

"Cross-Fitting and Averaging for Machine Learning Estimation of Heterogeneous Treatment Effects"

We investigate the finite-sample performance of sample splitting, cross- fitting and averaging for the estimation of the conditional average treatment effect. Recent meta-learners make use of machine learning to estimate nuisance functions and hence allow for fewer restrictions on the underlying structure. This includes splitting the data to reduce bias and averaging to restore efficiency. We employ a Monte Carlo study with different data generation processes and consider twelve different estimators that vary in sample splitting, cross-fitting and averaging procedures. We further use four different meta-learners: the doubly-robust-, R-, T- and X-learner. We find that the performance varies among the estimators. The best results among sample split estimators can be achieved when applying cross-fitting plus taking the median over multiple sample splitting iterations. Some meta-learners exhibit a high variance when the lasso is included. Excluding the lasso decreases the variance and leads to robust and at least competitive results.

"Does Tenure Make You Love Your Job? A Machine Learning Approach"

In this paper, we estimate heterogeneous treatment effects of having tenure on the perceived degree of six job satisfaction variables, namely, opportunities for advancement, intellectual challenge, level of responsibility, degree of independence,contribution to society and job Security. We use Survey of Doctoral Recipients data of year 2017 from the National Science Foundation (NSF) and focus on the PhD holders who work in academia. We explore the heterogeneous treatment effects(regarding gender, age etc.) using the conditional average treatment effect (CATE)and machine learning methods to extract group effects, defined as quantiles, from the CATE through a linear projection. The empirical findings support the hypothesis that tenure has heterogeneous causal effects on all satisfaction variables related to the job. We find the largest treatment effect on satisfaction with job security and opportunities for advancement. For certain groups, we even find small but negative effects from tenure which provide the evidence of heterogeneous treatment. We further conduct a classification analysis to provide insight into the average values of key features for the most and least affected. Here we find strong differences for gender, age, foreign-born and whether individuals have children or not. The findings show that new models in combination with machine learning reveal more complex effects rather than estimating the average treatment effect.

"Group Average Treatment Effects for Observational Studies"

he paper proposes an estimator to make inference on key features of heterogeneous treatment effects sorted by impact groups (GATES) for non-randomised experiments. Observational studies are standard in policy evaluation from labour markets, educational surveys, and other empirical studies. To control for a potential selection-bias we implement a doubly-robust estimator in the first stage. Keeping the flexibility to use any machine learning method to learn the conditional mean functions as well as the propensity score we also use machine learning methods to learn a function for the conditional average treatment effect. The group average treatment effect is then estimated via a parametric linear model to provide p-values and confidence intervals. The result is a best linear predictor for effect heterogeneity based on impact groups. Cross-splitting and averaging for each observation is a further extension to avoid biases introduced through sample splitting. The advantage

of the proposed method is a robust estimation of heterogeneous group treatment effects under mild assumptions, which is comparable with other models and thus keeps its flexibility in the choice of machine learning methods. At the same time, its ability to deliver interpretable results is ensured.

"Affordable Uplift: Supervised Randomization in Controlled Experiments"
Customer scoring models are the core of scalable direct marketing. Uplift models provide an estimate of the incremental benefit from a treatment that is used for operational decision-making. Training and monitoring of uplift models require experimental data. However, the collection of data under randomized treatment assignment is costly, since random targeting deviates from an established targeting policy. To increase the cost-efficiency of experimentation and facilitate frequent data collection and model training, we introduce supervised randomization. It is a novel approach that integrates existing scoring models into randomized trials to target relevant customers, while ensuring consistent estimates of treatment effects through correction for active sample selection. An empirical Monte Carlo study shows that data collection under supervised randomization is cost-efficient, while downstream uplift models perform competitively.

Daniel Jacob
Berlin, November 2, 2020