Data Mining Cup 2023

Statistics: Michel Lang, Jonas Rieger, Steffen Maletz

Computer Science: Emmanuel Müller, Simon Klüttermann

DMC 2023

- This is an on-site course
  - 12 participants from Statistics department and 10 from Computer Science
- Predictive modeling competition from the field of online marketing
  - http://www.data-mining-cup.de/
  - Training dataset + unlabeled test data for prediction.
  - Optimize against specified quality measure
- International competition
  - 2022: 78 Teams from 59 universities in 23 countries
  - 2021: 115/86/28, 2020: 162/126/35, 2019: 149/114/28, 2018: 193/148/47, 2017: 202/150/48
- (Successful) history of Dortmund statisticians
  - 2010: Second Place, 2011: First Place, 2013: First and Second Place
  - 2020: First and 6th Place (joint team with computer science department),
  - several top 10 occurrences
- Prize money (2000/1000/500 €)

Statistical Methods

- EDA (Explorative Data Analysis)
- Preprocessing (Imputation, ...)
- Resampling and Evaluation
- Discriminant Analysis
- Nearest Neighbours
- Trees and Forests
- Support Vector Machines
- Regularized Linear Models
- Gradient Boosting
- Neural Networks
- Hyperparameter optimization
- Feature Selection
- Feature Generation
- Ensembles and Stacking
- [...]

Software

- Version management using GitHub
- Visualization (interactive)
- data.table / SQL
- Parallel computing (local/cloud)
- Machine Learning frameworks
  - e.g. mlr3 in R or scikit-learn in Python
- Modern ML packages
  - e.g. ranger, xgboost, glmnet, sklearn
- Matrix as team chat for communication

Course Plan

- March 9 10:00 - 12:00, CDI 121: Workshop day I
  - we present exemplary machine learning workflows
  - we provide you an example dataset which you have to analyse
- April 4 & April 6, each 10:00 - 12:00, CDI 121: Workshop days II
  - you present your findings on the analysed dataset (and we provide feedback)
- TBA: Registration start DMC (2 team “leader” who register)
- TBA: Start of competition, release of data and task (GK Artificial Intelligence for Retail AG)
- During lecture period: Regular meetings (2 per week), **active participation**
  - Tuesday and Thursday, each 10:15 - 11:45, CDI 121
  - Statistics: short report (max. 10 pages) until TBA on EDA and first modeling approach
  - Computer Science: presentation on suitable models and techniques
- TBA: End of competition, upload of predictions for test data
- (TBA: Award ceremony)
- August 31: **Statistics: Final Report** (~ 25 pages)

Requirements

- Familiarity with data analysis tools like Python/sklearn, R or Julia

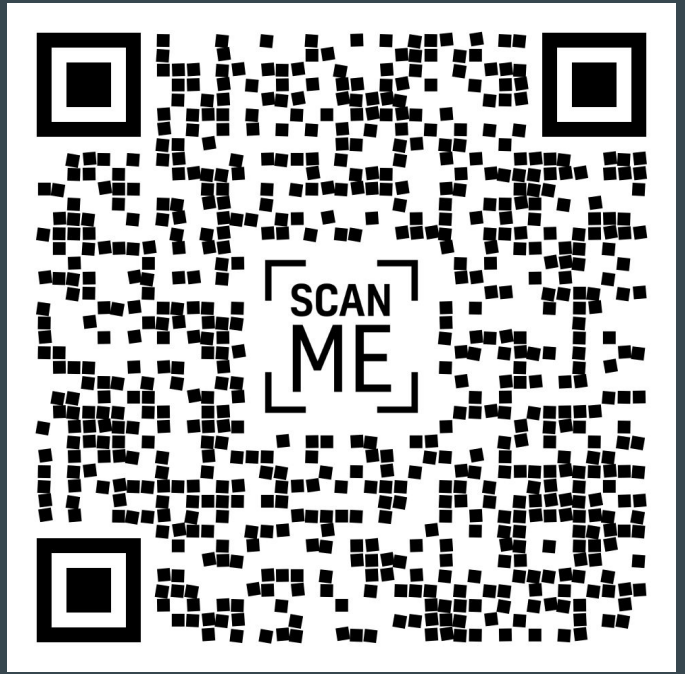
- Master Statistik: Fallstudien I (recommended)
- Master Econometrics: Minor Introductory Case Studies
- Master Data Science:
  - All requirement courses (Introductory Case Studies, ...) must have been passed
  - Advanced Statistical Learning is recommended to be passed

- Computer Science: Big Data Analytics (recommended), Mathematics Courses

Registration

- Statistik/Econometrics/Data Science:
  - starting now until February 26: https://umfragen.tu-dortmund.de/index.php/738348?lang=en
  - March 1: feedback whether your registration was successful

- Computer Science:
  - see respective registration system



Examination Statistics

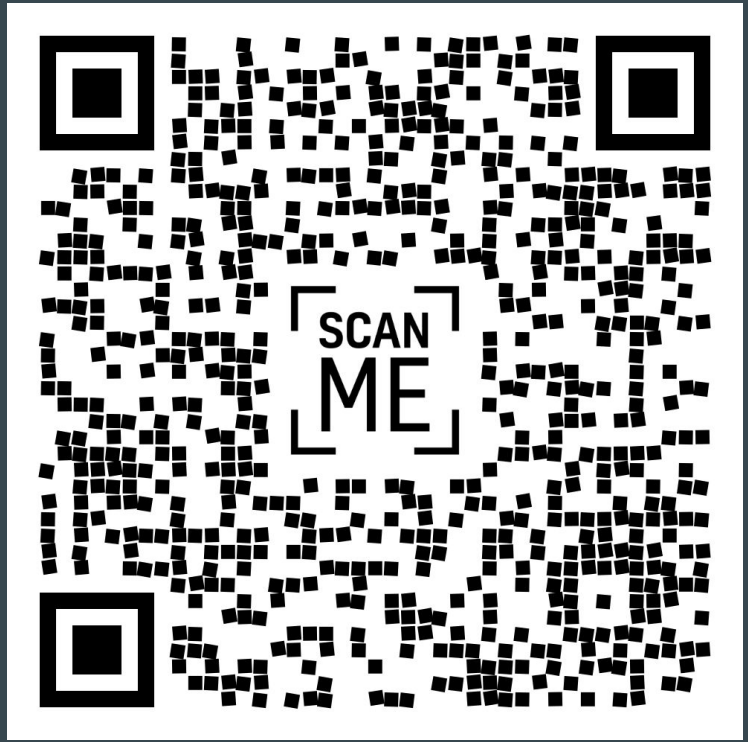
- short report (max. 10 pages) - deadline: TBA – no extension!
  - figures and tables count towards the page limit! Number of pages from the introduction until the last page of the conclusion is relevant for the page count; title page, contents page and bibliography do not count towards page limit
  - explanation of the given research question
  - scientific explanation of the presented method/concept
  - exploratory data analysis (EDA) of the given dataset
  - application of your method/concept to the dataset
  - structure of report (examplarily): 1 Introduction, 2 Problem Definition 3 Methods, 4 Application, 5 Conclusion/Discussion
  - take the rules as teached in case studies (Fallstudien I), Introductory Case Studies or similar... (or maybe thesis) as a guideline
- active participation in competition and discussions
- final report (~ 25 pages, we will announce specific formalia for this report at the end of the competition) - deadline: August 31 – no extension!

Examination Computer Science

- active participation in competition and discussions
  - initiative for open tasks
  - imagination for what could be useful tasks
  - take and fill necessary roles in team
  - think both in and beyond your team
- poster session
  - explanation of task, teams and your role in the DMC
  - outline how your team's process going from early to later solutions
  - explain team's contributions to the final solution

To Do

- Questions?



- Register (https://umfragen.tu-dortmund.de/index.php/738348?lang=en)
- If “accepted” as participant (you will get a mail by March 1):
    - keep yourself up to date using Matrix, we will share all information there
    - first task: determine 2 team “leader” for competition registration