Data Mining Cup 2022

Michel Lang, Jonas Rieger, Steffen Maletz

Computer Science: Emmanuel Müller, Benedikt Böing, Simon Klüttermann

DMC 2022

- We are planning the event as an on-site course.
- We will switch to an online course only if the situation requires it.
- 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
  - 2021: 115 teams from 86 universities in 28 countries
  - 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
- Slack for communication
- Zoom for potential remote meetings

Requirements

- 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
- Communication
- Teamwork

Course Plan

- February 16, 10am: Registration deadline
  - register at https://umfragen.tu-dortmund.de/index.php/224132?lang=en
  - max. 12 participants
- February 23, 10am: Remote Kickoff-Meeting, assignment of presentation topics
- March 21 - March 25: In-person "Bootcamp" week
  - **Student presentation on selected methods** (~ 20 min)
  - **Submission of short reports** (~ 10 pages): Application of presentation topic to exercise data set
  - Basics of Software Engineering
  - Basics of Machine Learning
- ~April: Start of the Data Mining Cup - release of data and task by the prudsys AG
- Regular in-person meetings (2-3 per week), **active participation**
- ~June: End of competition, upload of predictions for test data
- (~July: Award ceremony)
- End of August: **Final Report** (~ 25 pages)