Machine Learning for Imputation

A short overview over our study and what it brought/will bring

History

- ML started in Germany's official statistics at around 2014, very experimental, to make things done that no one had the time for before.
- Example: Classification of craft vs. non-craft enterprises (in the sense of our statistic law vs. in the sense of the chambers)
- Early success and good connections to academia provoked interest at higher levels in hierarchy
- 2018: Proof of concept machine learning (in general), installation of a central section for machine learning and imputation methods, begin of the study on ML methods for imputation

Current situation

- **5 people (from computer science, mathematics, economics, bioinformatics) for machine learning within Destatis**
- **working on**
  - **concrete projects (like text classification, imputation, transferring patters from one to another statistic) or**
  - **conceptual questions (new methods, quality aspects, cooperation)**
- **still support from hierarchy but also expectations to justify the staff costs for ML**
- **IT is still a problem (a long process to get money, a longer process to order, a much longer process to get access …)**

Simulation study

- **We knew before:**
  - **ML performs very well in classification and regression**
  - **trees perform well in imputation tasks (although it is sometimes needed to help them …)**
  - **there is a theoretical approach to imputation (Rubin & Little …) which is mostly for situations where the downstream task (e.g. the variable(s) of interest) are known at imputation time**
- **We also knew before:**
  - **just to make good predictions should not lead to good results because the stochastic element is ignored and by this we will artificially reduce the variance of a variable (i.e. also the estimated variance in a downstream task)**
  - **there are situations where we do not know what the downstream task is (e.g. when we deliver to Eurostat or publish tables online)**

Simulation study

- As you know: We found that (weighted) k-nearest-neighbor and random forests performed better than expected although we did not use any stochastic element.

- Interestingly, CBS found similar results independently from us.

- We heard from the CBS study via an upload on Statswiki ☺!

- Next steps were:
  - Contact with CBS: Planning a joint empirical study on these findings (with more official data sets and different structures of variables)
  - Presenting the results on several conferences and workshops
  - Looking for researchers that share their theoretical insights on this phenomenon with us

Simulation study

- **If the results are stable, the use of random forests would be much faster than traditional methods and than k-nearest-neighbors.**

- **Current status:**
  - **waiting for CBS**
  - **cooperation with the Technical University of Dortmund in order to find circumstances where one can prove or disprove that random forests do the imputation job well**
  - **bringing *missForest* into production (parallel to CANCEIS) in Destatis's new structure of earnings survey in order to be able to compare these two methods over the next years**

Hope for the future

- **to still have enough time for conceptual things like this study**
- **to not lose the support by Destatis's hierarchy**
- **to be able to extend the cooperation with universities (for this special question but also for questions on quality and on the compatibility of ML with complex survey designs)**