DESTATIS
Statistisches Bundesamt

November 16/17, 2020 | Florian Dumpert

The UNECE HLG-MOS Machine Learning Project: Report of the Editing & Imputation Group

Theses

- Machine Learning (ML) holds a great potential for statistical organisations.
- It can make the production of statistics more efficient by automating certain processes or assisting humans to carry out the process.

https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home

Pilot Studies

Editing

find missing and problematic data

correct incorrect data

Imputation

insert missing values

- United Kingdom: Editing of LCF survey data
- Italy: Validation of administrative data sources and some ideas and hints

extra talk

- Belgium: Imputation in the energy balance of Flanders

→ BigSurv 20

- Italy: Imputation of the variable “attained level of education”

extra talk

- Poland: Imputation of tourist expenditures and imputation of sports clubs
- Germany: Simulation study on machine learning for imputation

eu2020.de

Federal Statistical Office of Germany (Destatis)

Expectations

Editing:

- ML may **help to design rules** at the very beginning or **discover rules** that have only been “known” by intuition at first and trained in previous experience.
- ML could **learn from former editing results** which units (records or even cells) in a data set are problematic.
- ML may offer a valid and efficient **new instrument** for the not rule based perspective on editing.
- ML offers a **new approach to outlier detection.**

Imputation:

- ML may **improve prediction tasks** within already existing imputation schemes.
- ML may be **faster in doing imputation** compared to other methods once the model is learnt.

What we could observe

Editing:

- Learning from former editing results is possible: It is **possible to predict** whether a unit needs special attention.
- The **extraction of rules suffers** from the trade-off that good predictions are only achievable with very detailed (i. e. long and complex) rules.

Imputation:

- ML delivers **comparable** (compared to other methods) **results in a more automated way**.
- ML often produced **plausible predictions**.
- ML can produce **more timely statistics** by skipping some pre-treatment.
- ML can **reduce human intervention**.
- Projects with **time series** were successful.

What we have learnt

Editing:

- ML is **faster and more consistent** than manual editing.
- Training the ML model, however, is **expensive.**
- ML allows using huge amount of data with **much less a priori knowledge, hypotheses and data preparation**.

Imputation:

- ML is more powerful because of its property that **fewer assumptions** are needed.
- Usage of machine learning successfully in production is possible **only after a lot of (successful) experimentation** on the topic.
- Need to **shift the interest** to accuracy and timeliness of results rather than to the interpretation of the parameters.

Don’t forget

- Applying machine learning methods needs a bit more data science skills (programming, coding, training/testing principles) than using other methods.
- Programmers, statisticians, subject matter experts have to work together intensively.
- Machine learning and statistical methods have always only a serving role in the processes in official statistics. They can assist the subject matter experts and the management in their decisions.
- The usage of machine learning is only useful if it is better than the currently used baseline method and more simple statistical methods.

Conclusion (a preliminary overall assessment)

ML is suitable for editing and imputation in official statistics in principle.

Further investigations on the quality of the processes and the results have to be conducted.

However, a positive tendency can be stated.

Additional remark (citing Mark van der Loo)

ML might be useful for extracting/collecting information from trade magazines, newspapers, websites, financial reports, media, … for external validation of the data at hand.

Questions

- **Do you use ML for editing and/or imputation?**
- **Why?**
- **Do you make similar or completely different experiences?**
- **Are there best practise examples you can tell us about?**