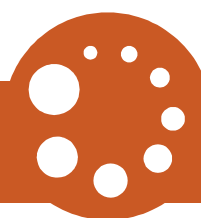




Plausi++: Automatic data validation to check the quality and reliability of administrative data and survey data

Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2019

Published by: Federal Statistical Office (FSO)
Editors: Dr. Christian Ruiz
Contents: Dr. Christine Ammann Tschopp, Elisabeth Kuhn, Laurent Inversin, Dr. Mehmet Aksözen, Stefan Rüber
Topic: 00 Statistical Basis and Overviews
Original text: German
Translation: FSO language services

Layout concept: Section DIAM
Downloads: www.statistics.ch
Copyright: FSO, Neuchâtel 2019
Reproduction with mention of source authorised (except for commercial purposes)



Pilot project Plausi++: summary

30 June 2019

Data validation is carried out to review the quality and reliability of administrative data and survey data. Data that appear to be either clearly incorrect or at least questionable are sent back to the provider with a request for correction or comment. Up to now, such data validation has been carried out either using automatic rules and thresholds or manually. In this project ("Plausi++") we developed a new type of data validation based on "Machine Learning" (ML) algorithms, which automatically identifies potential error patterns. We combined this with an automatic feedback mechanism, which can provide an explanation that is comprehensible to a human back to the data provider. This has allowed us to supplement the existing checks, to reduce the time needed and to improve data quality.

Results

Objective 1 ML algorithm ("predictive part"): Use of the ML algorithm considerably improved data quality: in a baseline comparison with existing data validation, Plausi++ **identified 29% more errors**. The ML algorithm has a **prediction accuracy of over 94%**.

Plausi++ **only** deals with **content errors**. Plausi++ **thus complements** the existing data validation, in which only 21% of rules were content-related and 79% **were of a formal nature** (e.g. code compliance, no blank data fields). **Plausi++ does not replace the existing data validation since symbiosis of the two systems covers both content-related and formal errors.**

Another indication of improved data quality is **the detection of systematic errors at aggregate level**. **Plausi++ identifies more systematic errors overall**. For example in one application, **twice as many** such systematic errors were found than with existing data validation.

There are two phases in "supervised ML". While the first phase takes a lot of time to train the algorithm, the second phase – the application phase – of **Plausi++** can be performed in **just a few minutes**. In addition, reducing the to-ing and fro-ing with data providers will most probably result in a time saving, but this can only be tested once it is implemented in the production environment.

Objective 2 Feedback mechanism ("explanatory part"): An automated feedback mechanism explains the result of the ML algorithm to a data provider in a way that a human can understand. Such a response indicating the potential problem variable (e.g. age or staff category) was fully implemented with one method. No other approaches have been tested yet.

Objective 3 Documentation: Detailed documentation was drawn up with a view to scalability so that other divisions and sections can understand and adapt it.

Objective 4 Integration in production: The pilot project assessed the feasibility of production integration within the limits possible. Needless to say, there are many unknowns, such as the definitive IT infrastructure solution and other elements that were not part of the pilot project. However, we have recorded our considerations and believe that Plausi++ can be effectively integrated into the production within the existing data validation process.



Conclusion

The pilot project achieved all four objectives and following a number of follow-up tests, is ready to be integrated in production. Integrating an ML algorithm in the strict sense was relatively simple, but the feedback mechanisms were more difficult. Detailed documentation is available with respect to the scalability of the ML algorithm.

Complementary analysis methods used:

- For the predictive part we used algorithms such as “gradient boosting machines” and “random forests”, achieving greater prediction accuracy with the former.
- In the explanatory part we used “*Ceteris Paribus*” plots and “Local Interpretable Model-agnostic Explanations” (LIME). We have currently set up the feedback mechanism using *Ceteris Paribus*.