Improving Data Validation using Machine Learning

Keywords: Machine Learning, Data Validation, Prediction, Explanation

1. INTRODUCTION

1.1. Aim

The aim of this project is to extend and speed up data validation at the Swiss Federal Statistical Office (FSO) by means of machine learning algorithms and to improve data quality.

1.2. Data validation

Statistical offices carry out data validation (DV) to check the quality and reliability of administrative data and survey data. Data that are likely to be incorrect are sent back to data suppliers with a correction request. Until now, such DV have mainly been carried out at two different levels: either through manual checks or automated processes using threshold values and logical tests. This process of two-way "plausibility checks" involves a great deal of work. In some cases, staff is required to manually check the data again, in other cases rules are applied that often require additional checks. This rule-based approach has developed from previous experience but is not necessarily exhaustive and always precise. Machine learning has the potential to ensure faster and more accurate checks.

1.3. Context

This project is one of the five (pilot) projects currently being developed in line with FSO's data innovation strategy (FSO, 2017) with the goal to augment and/or complement the existing basic official statistical production at the FSO.

2. METHODS

2.1. Using innovative new ways of machine learning to find alleged mistakes

This approach would rely on a machine learning algorithm using historical data first. Based on previous analysis, a target variable can be defined that should be able to be predicted by the algorithm. Only then can the algorithm be used for the prediction. As the final stage, the predicted and actual values of the target variables are compared and the predictive accuracy can be evaluated.

2.2. Using innovative methods to explain the alleged mistakes

In the second part of the project, a feedback mechanism is used to send an automatic explanation to data suppliers. This is necessary, as it is impossible to combine high-prediction performance and interpretability with the same algorithm. Thus while we achieved a strong prediction in the first part, the same algorithm can not serve for

explanation. In the second part we thus build a feedback mechanism, a 'local explanation', to open the "black-box".

Initial ideas of this project were also presented by Ruiz (2018).

3. CONCLUSIONS

It will be the first time that we will publicly present the initial results of the ongoing (pilot) project. The results of the first part of the project are convincing and show us that those cases that have a high-predicted posterior probability to be wrong seem indeed wrong through manual checks. Further, a comparison of the new and the previous DV methods indicates an improvement of the data quality. The second part related to the feedback mechanism is still more experimental and we found different possible alternatives of generating 'local explanations' to provide an automated feedback.

.

REFERENCES

[1] FSO (2017). Data Innovation Strategy, (https://www.bfs.admin.ch/bfs/en/home/news/whats-new.assetdetail.3862240.html and https://www.experimental.bfs.admin.ch/en/).

[2] Ruiz, C. (2018). Improving Data Validation using Machine Learning, UNECE Workshop on Statistical Data Editing Working Paper (http://www.unece.org/index.php?id=47802).