Machine learning for Data Editing NSI

Unece ML project

ML for editing:

ideas, hints and some experiences

Fabiana Rocci

Istat

What is Editing&Imputation

- *Non sampling errors:* any error in collected data
  we treat the case when the observed value is different from the actual "true" value

- Important consideration:
  only sometimes it is possible to regard a data as *not correct*

  most of time it is only possible to regard it as *suspicious*

  usually further evaluation is needed to understand whether:
  i. a sopsicious data is an *actual* non sampling error, hence to be changed («imputed») to make it coherent with the given rule (still…it is reccomended no to call it 'correct')
  ii. instead it is *anomalous* but it is not an error (probably an *outlier*)

What is Editing&Imputation

- Editing: it can be considered as a problem to define the right function to *label* records between correct and erroneous/suspicious

- Imputation: methods to treat data found to be not correct or missins data

About Editing

- Editing defined as “indentifying suspicious data”
- It has been underlined that across the NSO this is the less investigated area
- During the first meeting we had a lot of discussion how ML could help in editing

ML project: starting review and PoC 1/3

- During the sprint meeting in London from an overview of literature and experiences in NSI offices of ML, has resulted

  - very few application have been done so far in the editing part regarded as the part **to detect** data containing *non sampling errors*
  - most applications are related to the imputation for missing data

ML project: starting review and PoC 2/3

- At start only one PoC was presented:

Edit and Imputation of LCF survey data (*Claus Sthamer, Office for National Statistics, UK)*

"So far, there is only manual detection of spurious records. The goal was to replace the need for manual detection by learning a supervised model from former editing steps."

- After some months another PoC joined the project:

The new Statistical register on economic variables of the Public Administration: the expected results was that sharing ideas could help in driving us in the design of a completely new E&I scheme design
(*Roberta Varriale, Fabiana Rocci, Istat, IT*)

"No legacy system, the task is new. Edit rules are of the main focus, but there are also investigations whether the application of machine learning can add value to the traditional editing process. "

ML project: starting review and PoC 3/3

- *we were left with an homework:* to analyse the potential use of ML for the editing part

- Some ideas during the sprint meeting were suggested by the E&I group

- In this view, a cooperation among ONS (PoC on Editing), Destatis and Istat started to study the potential use of ML for editing, meant as only the functions to detect the suspicious/errors data

E&I reference scheme: criteria 1/2

A. Schemes for detecting errors usually starts from classifying the **types of errors by source and effect on data:**

i. the potential *source and nature*:
mainly divided in **systematic and random**

ii. degree of danger on the final estimates: **influential or not** (critical or not)

E&I reference scheme: criteria 2/2

B. For each type of errors **methods are defined**, they can rely on :

i. Edit rules: An edit rule is a restriction to the values of one or more data items that identifies missing, invalid or inconsistent values.
Edits are often distinguished between *fatal* or *hard* edits and *query* or *soft* edits, depending on whether they identify errors with certainty or not

ii. Score or outlier detection methods: assess the plausibility and influence of the values of a unit with regard to the aggregate estimate. They are functions defined to release a score to measure the plausibility.
Sometimes also outlier detection methods are used, they can help in guiding to put more attention on a set of units that sound to be too much different form the rest of the data distribution.

Type of errors, methods and flow:

A. Systematic errors:

Domain / Obvious and other systematic errors: Check of structural informative objects defining the target population and the variables/ Obvious errors are the ones easily detectable and treatable. The remaining Systematic errors, can be reported consistently across units

Methods: based on rules, both hard and soft

B. Influential errors: Influential errors are errors in values of variables that have a significant influence on publication target statistics for those variables.

Methods: based on score or outlier detection function

C. Completely at random not influential errors Random errors are errors that are not caused by a systematic reason, but by accident and they do not result to be influential.

Methods: based on rules, both hard and soft

ML for editing: about edit rules

There is a **major reflection** about the definition of the **soft /plausibility rules**:

they come from hypothesis about the relationship between the data and some other characteristic, typically between several variables of the same record or a plausible relationship with regard some characteristic feature of the group they belong (typically: mean)

But **those rules should not be/could not be taken for grant**, neither across the several units nor along the time on the same units**.**

It can cause high cost to invest resource in analysing the result of *labeling* by those rules

ML possible value added: some hints 1/1

- Supervised ML: it can be thought to be used when a *training set* of data where data are *labelled* already between 'plausible' and 'not plausible' is available

  Hint: it is possible to approximate the function labelling the data as correct or not and to apply it to new data. It can help in reducing the uncertainty about the plausibility rules

  Improvements can release a complete automatic new process when:

  - ✓ the estimated function is demonstrated to catch all the labelled data perfectly…that is such a rare event!
    BUT: our task/hope is that it could help in describing in a better way the rules or at least to reduce the number of observation to be regarded as suspicious
  - ✓ rules do not change over time

ML possible value added: some hints 1/2

- Unsupervised ML: when data that are unlabelled and/or it is believed that data are characterized by completely hidden patterns of errors. In this sense, unsupervised ML has the potential to identify relationships among variables.

Hint:

i. an unsupervised ML approach can help to classify the data based on the patterns or clusters

two phase ML: this process adds labels to the data so that it becomes supervised. Therefore, unsupervised learning can be used as the first step before passing the data to a supervised learning process.

ii. when the problem is to identify groups of data, this could lead to outlier detection or to pattern detection of errors as strange data with regard the rest of the distribution: in this regards, the wish is to learn the inherent structure of our data without using explicitly-provided labels.

What has been done: 1/2

- ONS PoC

  method: randomForest

  Supervised method to label records: it could be replied automatically on a new set of data from the same population.
  Mostly the aim was for random erros.

From the E&I WP report :

"where the aim has been to analyse the capacity of the use of ML to increase the automation of the editing phase as much as possible, i. e. to reduce interactive editing in favour of automation, showed
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."

What has been done: 2/2

- Istat PoC:

  method: classificationTree

  Supervised method to extract rules for influential data: it could help in defining the rules that had been used but defined 'implicitly' by the thematic experts

  The project is still under analysis, first results:

- Among more than 100 variables, a small group have been discovered to have major impact on the final prediction of influential data
- Going region by region: different group of variables impact the predictions in different ways
- These hints started to drive the design of editing procedure. A further application of the classificationTree is foreseen
- In a way, we are moved from "to predict" **towards** "to give support to decision making" in a validation process
- Identifying hidden pattern in data to save time and resource for the interactive editing

E&I reference scheme : guidelines from

- Edimbus - Recommended Practices for Editing and Imputation Cross-Sectional Business Surveys, EDIMBUS project report, 2007, Eurostat

- Memobust – handbook for every phase of a survey on businesses, a complete chapter about E&I in every regard is provided
  Handbook on Methodology of Modern Business Statistics, CROS-portal, 2014, Eurostat

- General Statistical Data Editing Model: it supplies a general scheme for designing an E&I process, defined in functions, methods and process step
  GSDEM, Unece, 2019

- ML project reports available at:
  https://statswiki.unece.org/display/MLP/Project+documents
  https://statswiki.unece.org/display/MLP/Working+documents