Editing of Social Survey Data with Machine Learning
A journey from PoC to Implementation

Organisation: ONS (Office for National Statistics) - UK
Author: Claus Sthamer
Date: 15/10/2020

This paper is about a Proof of Concept (PoC) investigation into the Editing and Imputation theme. It was started as a contribution to the UNECE ML project to investigate if Machine learning (ML) can be applied to the identification of suspicious, implausible or erroneous personal income data records of the Living Cost and Food (LCF) survey that need clerical error correction.

Introduction

The UNECE Machine Learning project was recommended by the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) – Blue Sky Thinking network in the autumn of 2018, approved early 2019 and launched in March 2019. The objective of the project is to advance the research, development and application of ML techniques to add value or to make the production of official statistics better, where 'better' could be defined as:

- Cheaper
- Enabling faster releases of data
- Having more consistent data
- Providing alternative data sources

The project participants agreed to investigate these three themes in Work Package 1 (WP1):

- Classification and Coding (C&C)
- Editing and Imputation (E&I)
- Imagery

Here we share some comments about this project from participants of the work package group:

"With rapidly growing interest in the use of machine learning for official statistics but with limited experience with concrete applications, there was a great need for a common platform where experts in national statistics offices could test their ideas and exchange experiences. National statistics offices work on similar type of problems and operate with similar business constraints, so they can benefit from developing shared understanding. From its inception, active participation in the machine learning project has demonstrated the value of this collaboration for members. Sharing codes and methods quickly spreads knowledge and enhances capabilities within the project team and the identification of common challenges led to discussions about common solutions. The growth of the project team from 20 people to 120 people within one and a half years shows the timeliness and usefulness of the initiative which is fully in line with the core mission of the UNECE HLG-MOS - work collaboratively to identify trends, threats, and opportunities in modernising statistical organisations." **InKyung Choi - UNECE**

"The last 18 months have proven that the use of Machine Learning (ML) for official statistics is growing exponentially. When we started this project, we had limited understanding what the importance would be to produce faster, better and often new data sets. New ways of working have been the key driver to inform everyone over the last 8 months with the very important COVID-19 output. From the first sprint hosted by the Office for National Statistics (ONS) in May 2019 we agreed that Machine Learning could play a key part in the different themes of statistical output. We had to address this through allocation of various POCs to the very unique themes of C&C, E&I and Imagery, where coding activities are prime candidates for ML algorithms. The need to automate Editing is vast and could replace time consuming tables and error prone spreadsheets, finishing with Imagery were the use of various satellite data will speed up processing of data enormously. All this can only be achieved through shared and collaborative working." **Eric Deeben – ONS Data Science Campus – UNECE ML Work Package 1 Lead**

Collaboration

This UNECE ML project has encouraged collaboration between participating organisations and members. We have seen examples of code sharing to get members very quickly up-and-running with good prediction results on their own data. This has enabled and kick-started organisations' ML journey which is one of the aims of this project. The Office for National Statistics (ONS) collaborated with DESTATIS, the German central statistics office and with Istat, the Italian statistics office. Detailed discussions on our respective Proof of Concepts (PoC) have increased understanding and given us ideas to progress.

"The theme on E&I presents a summary of the activities and experiences that have taken place across the NSIs, in order to identify the value added and to identify further possible developments in this field.

E&I addresses all the activities that are run along a statistical process to identify 'suspicious' data and to 'impute' them. There is a need for NSIs to identify and deal with suspicious and missing values in datasets. There are several ways to do this. During the first ML project meeting, all the PoCs for the E&I theme were presented. Discussion showed that some experiences using ML were already under way. Nevertheless, it was observed that most of the experiences were about methods to 'impute'. On the other hand, methods aimed at the identification of suspicious data were much less investigated. In this view, the ONS was the only one presenting on the Editing part and the first meeting was a chance to exchange some ideas about the extent to which ML could help in the specific field of editing related only to the identifying part. The collaboration started with the ONS working directly on the PoC, DESTATIS, who coordinated the group, and ISTAT who were also interested in the same theme.

The starting point was to consider the statistical methods to "detect" suspicious data as a problem of "classifying" data between being coherent or not coherent, to be treated more carefully. It was thought that a supervised ML model could learn from former editing results which units (records or even cells) in a data set are problematic.

To test, afterwards, whether such a model is sufficiently interpretable, rules can be extracted from it that represent one possible way to classify a unit as "plausible" or "not plausible". The collaboration started to exchange ideas and experience to which extent ML (and model-based approaches in general) may offer a valid and efficient new instrument to gain/achieve/build a new perspective on editing.

Several meetings were set up to share the experience gained by the ONS. This shared experience and discussions lead to new ideas to make it possible for ISTAT to start a new PoC as well. This led towards the idea of using ML to support the design of edit rules in a new process." **Fabiana Rocci - ISTAT**

"*How can official statistics make its processes more effective and/or more efficient? Based on this central question, it is almost impossible not to encounter machine learning these days. A revolution is promised, sometimes erroneously under the label "artificial intelligence": The computer will fix it; human work will play a subordinate role in the future. However, can this promise really be kept? For the field of official statistics, the UNECE Machine learning project should answer this question in terms of "if" and "how". Questions of an ethical, legal and social nature were not examined in this project.*

Relevant aspects of "whether" and "how" include on the one hand the concrete area in which machine learning is to be used (e.g., for the assignment of statistical units to predefined or still to be defined classes such as economic sectors or occupational groups; for the identification of suspicious, unusual data that may require correction; for the estimation of data that has not been collected, has not been reported or requires correction; for the handling of new types of input data such as satellite images, etc.), but also questions about quality aspects and the concrete integration into existing (IT) processes.

It was shown that comparable questions and experiences exist in the NSIs involved in the UNECE Machine learning project. The open exchange of experience and solutions on these issues offers benefit for the work in the statistical offices. Moreover, the co-operation provided insights into concrete projects, delivered application examples beyond the borders of the respective offices and, due to the trustful co-operation, allowed for in-depth discussions on methodology, implementation, and coordination within the respective NSI as well as with external *stakeholders and co-operation partners. All in all this is a highly rewarding project clarifying the view on machine learning in official statistics and showing that human work will still play an important role in the future."*

Florian Dumpert - DESTATIS

Sharing Knowledge

Another aim is for members and interested parties to share code and knowledge between them. Most members have PoCs and some more advanced ML solutions that have been operationalised with code, tutorials and other materials available on GitHub repositories.

Background

Three ONS social surveys, the LCF, Survey of Living Conditions (SLC) and the Wealth and Asset Survey (WAS) will be combined to form the Household Financial Survey (HFS).

These surveys, as they are now, have the following numbers of cooperating households and survey specific themes:

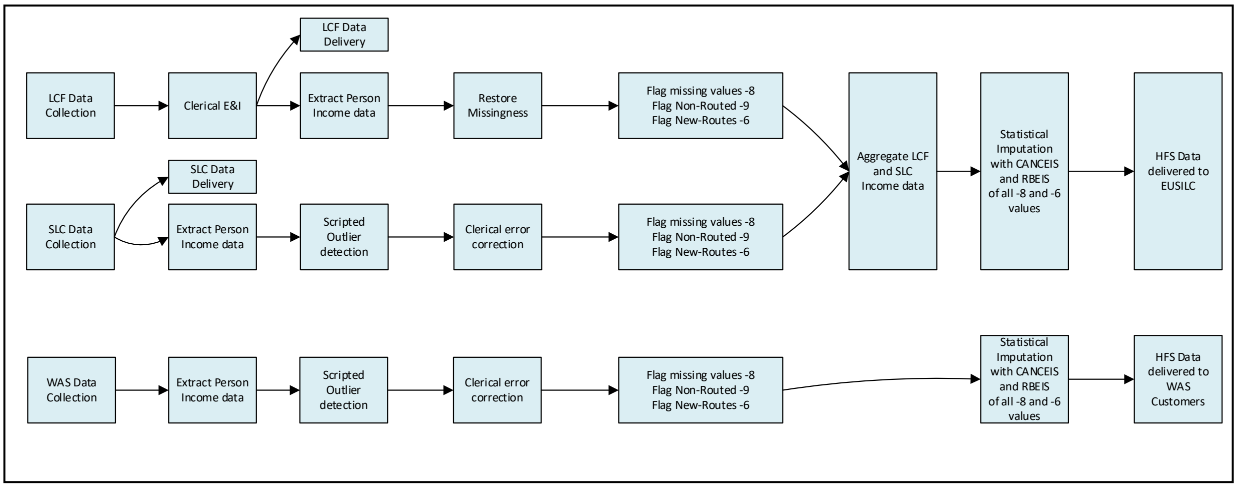
- LCF – 5,000 – Expenditure, Food & Nutrition
- SLC – 12,000 – Living Conditions
- WAS – 10,000 – Wealth & Assets

The themes will be retained for subsamples of the new HFS survey.

The existing survey pipelines are shown in Diagram 1 with their individual Editing processes:

- LCF – Clerical Editing and Imputation is carried out for the entire household record by a team within the ONS.
- SLC & WAS income data:
  - Editing - Scripted detection of outliers, followed by clerical value correction for these identified cases by the Validation team based at the ONS.
  - Imputation of missing data with CANCEIS and RBEIS.

Diagram 1 – Survey Pipelines


LCF Data Delivery
LCF Data Collection
Clerical E&I
Extract Person Income data
Restore Missingness
Flag missing values -8
Flag Non-Routed -9
Flag New-Routes -6
Aggregate LCF and SLC Income data
Statistical Imputation with CANCEIS and RBEIS of all -8 and -6 values
HFS Data delivered to EUSILC
SLC Data Delivery
SLC Data Collection
Extract Person Income data
Scripted Outlier detection
Clerical error correction
Flag missing values -8
Flag Non-Routed -9
Flag New-Routes -6
WAS Data Collection
Extract Person Income data
Scripted Outlier detection
Clerical error correction
Flag missing values -8
Flag Non-Routed -9
Flag New-Routes -6
Statistical Imputation with CANCEIS and RBEIS of all -8 and -6 values
HFS Data delivered to WAS Customers

The aim is to build a ML solution from this PoC for all HFS survey personal income data to predict the data records that need clerical error correction. The income block of the LCF and SLC surveys has already been harmonised (WAS only partially), to have identical income questions and data features.

This PoC was carried out on LCF data as they are intensively clerically edited with errors corrected and missing values being manually imputed. Because of this, they form a "Golden Standard" data set that can be used to label training data for Machine Learning. The 2018 Quarter 3 (2018Q3) data were used as training data and the 2018 Quarter 2 (2018Q2) data were used as test data.

Data Preparation

Various data preparation techniques were used to increase model performance:

1. From the 2000-person level features (survey variables) contained in a household record, 91 numeric and categorical features were selected from these areas:
    a. Income and tax
    b. Education
    c. Family situation
    d. Income and tax of job and secondary job
    e. Happiness and wellness
    f. Affordability of hobbies, clothes and shoes

2. Even though there are only numerical or categorical features used, not-numeric-values can be present due to Don't Know answers, refusals and not routed to. These are replaced with -1.
3. One-Hot-Encoding (OHE) of the categorical features.
   All possible values of a categorical feature must be represented as an OHE feature. The OHE process creates for each possible value of a categorical feature a new feature. Only the corresponding new feature will have a value of 1 if the original feature had that categorical value, all other features will have the value 0 for that record.
4. Normalisation of Net Pay and Gross Pay into annual amounts.
5. Aggregate 4 quality of life features (Satisfaction, Worth, Happy, Anxiety) into a new feature called Wellbeing
6. A Change Vector was calculated to label the records if there was a Change or No-Change of the data during the clerical Editing and Imputation process.

The Change vector features and frequencies for the training and test data are shown in Tables 1 and 2. The features used as predictors for the change vectors are the ones mostly changed during the clerical editing process. A 10% change threshold was used to eliminate all small changes.

Table 1 – Change vector features and Change Frequencies for 2018Q2

Feature	Description	Change Frequency	Change Frequency at 10% Threshold
NetNorm	Annual amount of net income (after deductions)	89	80
IncTax	Income Tax	270	268
NIns	National Insurance paid over given period	285	283
GrossNorm	Annual amount of gross pay (before deductions)	344	231
DedPenAm	Deduction for pension or superannuation	113	112

Table 2 – Change vector features and Change Frequencies for 2018Q3

Feature	Description	Change Frequency	Change Frequency at 10% Threshold
NetNorm	Annual amount of net income (after deductions)	52	49
IncTax	Income Tax	247	246
NIns	National Insurance paid over given period	275	273
GrossNorm	Annual amount of gross pay (before deductions)	319	207
DedPenAm	Deduction for pension or superannuation	93	93

The Random Forest algorithm from the Python sklearn library was used to classify the data into two classes: Change and No-Change.

The following Hyperparameters were used:

```
RandomForestClassifier(bootstrap = True,
                class_weight = 'balanced_subsample',
                criterion = 'gini',
                max_depth = 40,
                max_features = 'sqrt',
                max_leaf_nodes = 400,
                min_samples_leaf = 5,
                n_estimators = 1000,
                n_jobs = -1)
```

The hardware used to develop this PoC was:

- ThinkPad T490
- Intel Core i5-8365U
- 1.60GHz
- 8 GB RAM
- 256 Gb SSD

Results

The prediction results at the prediction thresholds are shown in Table 3:

Table 3 – Prediction Results

Prediction Threshold	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Recall	97.4%	93.5%	90.7%	85.5%	80.6%	77.5%	74.4%	71.8%	68.5%	65.9%	63.6%
Precision	38.1%	43.6%	53.6%	64.3%	74.6%	85.7%	92.0%	95.9%	98.1%	98.5	99.2
F1-Score	54.8	59.5	67.4	73.4	77.5	81.4	82.3	82.1	80.7	78.9	77.5
TP	377	362	351	331	312	300	288	278	265	255	246
FP	612	468	304	184	106	50	25	12	5	4	2

This PoC has shown that ML can be used to predict personal income data records that need clerical editing very well. The question "What is good enough?" must be answered to pick the prediction threshold to set the two embattled prediction quality indicators: Precision and Recall.

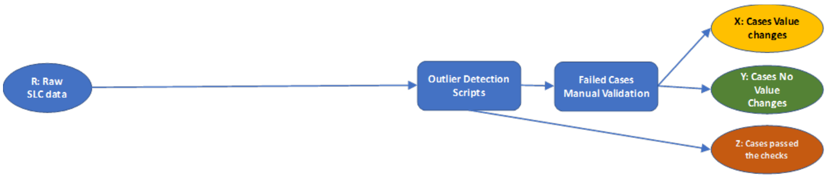
The journey towards Implementation

This PoC has also shown that with ML the number of LCF records that must be inspected, but then do not need any value changes applied, can be vastly reduced. It is hoped that this can also be achieved for SLC data and can, therefore, be a solution to the HFS editing dilemma. It is hoped that this method will become part of a unified editing pipeline for the 3 surveys, dramatically reducing the manual editing required.

The investigation into this is now under way. The existing editing pipeline for SLC data (see Diagram 2) uses outlier detection scripts that list cases and details about features and values that need corrections during the manual validation process.

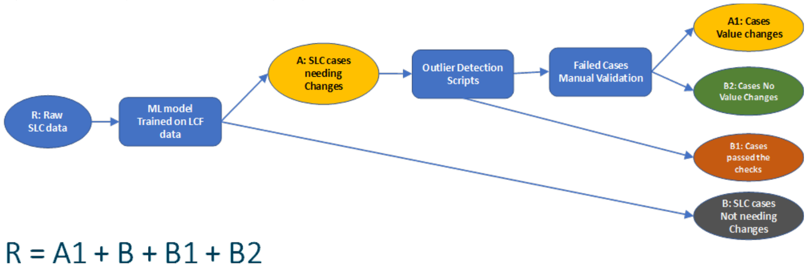
But it also lists cases that end up not having any values changed. This group of cases, sub-population Y, poses a big burden on the validation team that carries out this work. The aim is to minimise this burden while preserving the detection of cases that require data changes.

Diagram 2 – legacy SLC Editing Pipeline


R: Raw SLC data
Outlier Detection Scripts
Failed Cases Manual Validation
X: Cases Value changes
Y: Cases No Value Changes
Z: Cases passed the checks

The proposed SLC editing pipeline (see Diagram 3) incorporates ML into the legacy system. The ML model will be trained on LCF data to predict SLC data that requires human intervention. This transfer learning approach is expected to work well as the income data and question block has been harmonised between the LCF and SLC surveys.

Diagram 3 – Proposed SLC Editing Pipeline


R: Raw SLC data
ML model Trained on LCF data
A: SLC cases needing Changes
Outlier Detection Scripts
Failed Cases Manual Validation
A1: Cases Value changes
B2: Cases No Value Changes
B1: Cases passed the checks
B: SLC cases Not needing Changes

$R = A1 + B + B1 + B2$

Only the predicted cases classified as Change will then be fed into the existing outlier detection scripts. The objective is to minimise the number of cases in:

- Sub-population B1 in Diagram 3 that passed the scripts. These should have already been classified as No-Change by the ML model
- Sub-population B2 in Diagram 3 that failed the scripts but are then found not to have any value changes applied to them during the validation process. The expectation is that these have also been classified by the ML model as No-Change to avoid them to be fed into the SLC scripted outlier detection system.

The scripts can then hopefully identify all, or at least most remaining cases as sub-population A1, as cases that fail the scripts and need value changes during validation. One very important benefit of this pipeline is that the ML acts as a filter and can be switched on and off should conditions require this. The already existing process of manual validation could then be carried out regardless.

Value added by ML

From the results of the PoC it is expected that ML can add value to the HFS survey by:

- Making data available sooner for statistical analyses and publication.
- Reducing the manual data validation burden by filtering those cases out.
- Offering a common editing solution to all 3 surveys that are part of the HFS survey.
- Avoiding a 5-fold increase in clerical editing resource if the LCF method was applied to all of the HFS data.

Implementation Challenges

Some implementation challenges must be addressed and resolved:

- Model drift and monitoring
- Training data sets that can be labelled have to be re-built when survey questions, tax or benefit rules change
- ML explanations must be built into the solution to satisfy requirements set by the UK Information Commissioner's Office (ICO)