Editing of LCF (Living Cost and Food) Survey Income data with Machine Learning

Organisation: ONS (Office for National Statistics) - UK

Author(s): Claus Sthamer

Date: 08/06/2020

Version: 6.0

1. Background and why and how this study was initiated

This pilot study investigates if Machine learning can be applied to the identification of suspicious looking personal income data records of the LCF (Living Cost and Food) survey that need clerical error correction.

The LCF survey will be combined with the SLC (Survey of Living Conditions) and the WAS (Wealth and Asset Survey) to form the HFS (Household Financial Survey).

The aim is to build a ML solution from this pilot for the HFS survey to predict the personal income data records that need clerical error correction.

These 3 surveys, as they are now, have these numbers of cooperating households:

- LCF – 5000
- SLC – 12000
- WAS - 1000

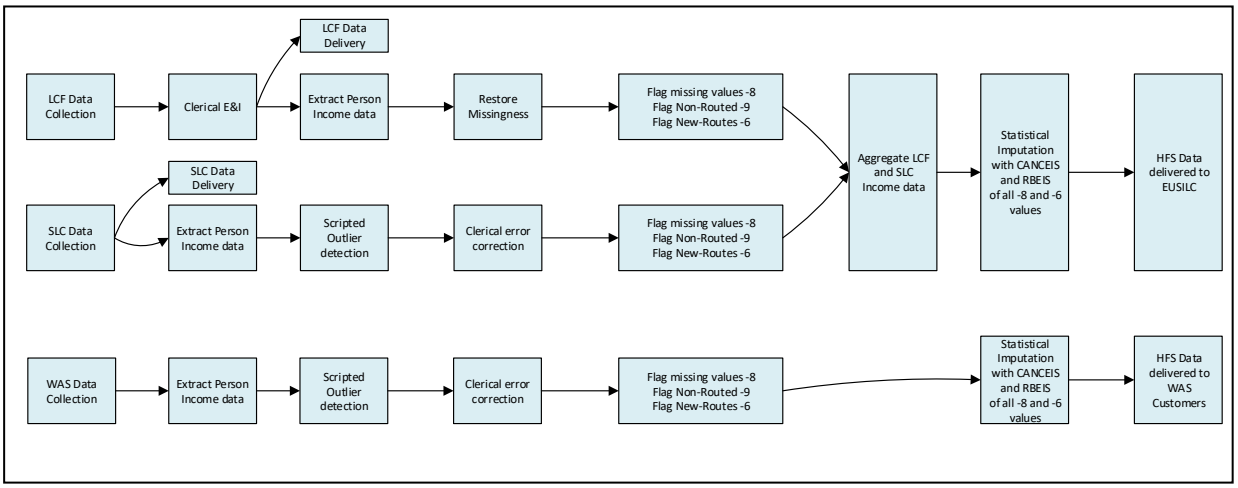
The survey specific themes:

- LCF – Expenditure, Food & Nutrition
- SLC – Living Conditions
- WAS – Wealth & Assets

will be retained for subsamples of the new HFS survey.

The income block of the LCF and SLC surveys have been harmonised, they have identical income questions. But the WAS income block is still in the process of being harmonised.

Picture 1 – Survey Pipeline



The existing survey pipeline is shown in Picture 1 with their individual Editing processes:

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

The LCF comprehensive clerical editing & imputation process is deemed to be over-editing and too time consuming. An experiment where LCF income data was put through the SLC scripted outlier detection process resulted in only 10% of the changes compared to the ones made by the LCF clerical editing & imputation team. This raises doubts about the accuracy of the SLC and WAS scripted outlier detection systems and suggests that the SLC and WAS data are under-edited.

As the LCF clerically edited data are the closest we have to a "Golden Data Set" or a ground truth and the fact that the income statistics conducted on LCF and SLC data differ, suggests not only that to carry out this pilot study on LCF data is the right approach, but also that SLC and WAS need some "fixing".

To allow for the survey merger to go ahead and preserve a similar data quality to that achieved with the LCF method, a new common process for editing and error correction has to be found that can deal with the vastly increased data volume compared to the LCF survey.

The scope of this pilot project is limited to LCF income data only, because the raw survey data and the clerically edited and imputed data are both available for this survey.

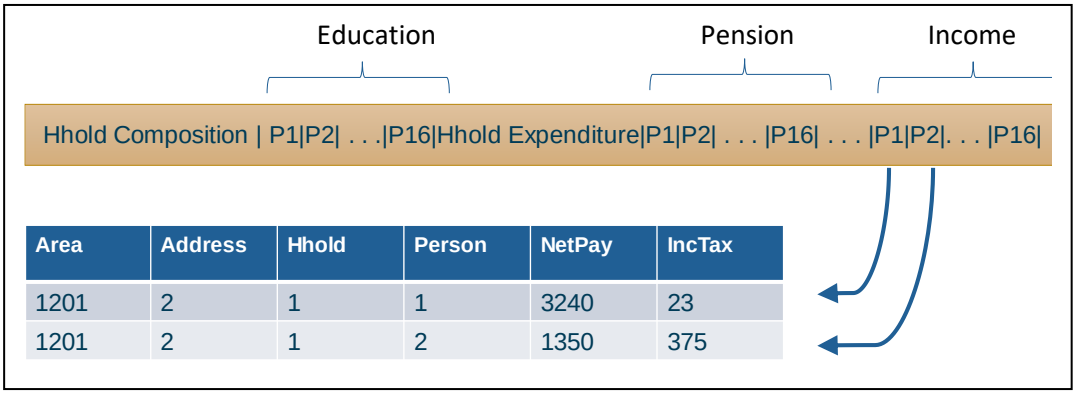
2. Data

2.1 Input Data

Even though the long term aim is to solve the Editing challenge for the HFS survey, this pilot study uses LCF survey data for the simple reason that both edited and imputed data and the raw survey data are available. This allows for changes of the data made during the clerical editing and imputation process to be labelled and used for supervised machine learning. The clerically edited and imputed LCF data are not checked by another team of experts for errors and these data form therefore not the Golden Standard that would be desirable, but they are the 'best' that is available. The LCF survey team has estimated that 80 % of all the changes made during the clerical editing and imputation process are correct. This level of accuracy has to be preserved.

Social Survey data at ONS is collected with the Blaise system, a computer-assisted interviewing (CAI) system and survey processing tool developed by CBS (Statistics Netherlands).
Mainly face-to-face interviews (CAPI – Computer Assisted Personal Interviewing) with some surveys or components of them using telephone interviewing (CATI – Computer Assisted Telephone Interviewing) are carried out.

Picture 2 – Household data Record



The structure of a household record is shown in Picture 2. There are topic specific questionnaire and data blocks. Some of these blocks are arrays for the up to 16 household members. The income, pension and education blocks are shown in Picture 2.
Income data and some other data from each member of a household is extracted to form the LCF personal income data set.

These personal income data sets were extracted for the raw data, these are the data received by the office from the interviewers and for the clerically edited and imputed data for quarter 2 (8Q2) and 3 (8Q3) from 2018, resulting in 4 personal income data sets as csv (Comma Separated Values) files.

It was decided to use the 8Q2 LCF data as the Test Data to be able to compare the 8Q2 ML prediction result with the result from the experiment when 8Q2 LCF data were put through the SLC system. This comparison has not been done when this report was completed. The Training Data are the 8Q3 data with 3059 Records and the Test Data are the 8Q2 data with 2912 Records.

2.2 Data Preparation

Various data preparation techniques were used as this pilot progressed 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.
   It is possible that some options of the categorical features might only be used in the 8Q2 records, but not in any one of the 8Q3 records or vice versa. For this reason, all possible values of a categorical feature have to be represented as an OHE feature in both data sets. Otherwise, the training and test data sets might not match in the number and names of features. One example is the NetPd feature, there are 15 options the respondent can choose from to indicate what period of time is covered with the last pay received. Option 8 is used to state that pay is received 8 times a year. The OHE process creates the feature NetPd_8 if there is at least one record in the data set where that option was selected. However, if no respondent has picked this option in the 8Q2 data set, NetPd_8 will not be created. But if at least one has done so in 8Q3, NetPd_8 will be created and there will be a mismatch in the features between the training and test data. The prediction method will fail.
4. Normalisation of Net Pay and Gross Pay into annual amounts.
   These features are paired with their time period, e.g. the value given for Net Pay (NetPay) has a feature that describes the period over which Net Pay was earned (NetPd). A respondent might have said that Net Pay is £1800 (Pounds Sterling) over a period of two calendar months. The annualised value will be 6 x £1800 = £10800
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,

Table 1 – Change vector features and Change Frequencies for 8Q2

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 8Q3

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

In early experiments, a binary change vector was made, where a Change in any of 25 selected features resulted in a '1' and a No-Change in a '0'. These 25 features were based on the 'Editing and Imputation Instructions for Income', used by the clerical team and are their target for error correction and insertion of missing values.

Experiments showed that by removing 20 of those features from that list with low change frequencies resulted in better predictions of the ML algorithm, it reduced the noise for the ML algorithm and allowed it to detect the more prevalent data characteristics. For example, the feature SeTaxAmt (Self-employment total tax) was only changed in 6 of the 2912 test data records. Errors in this and those other 19 features was very rarely predicted by the algorithm.

However, for the 5 remaining features, even the smallest of changes made to the data by the clerical editing team resulted into the label Change, e.g. the value of IncTax might have been changed by as little as £0.01.

We then used a relative change vector, where the percentage change was calculated. A relative change threshold of 10% was found to give best prediction results. Any changes of 10% and above were then counted as a Change and written as a '1' into the change vector and changes below 10% as a No-Change and a '0' was written into the change vector.

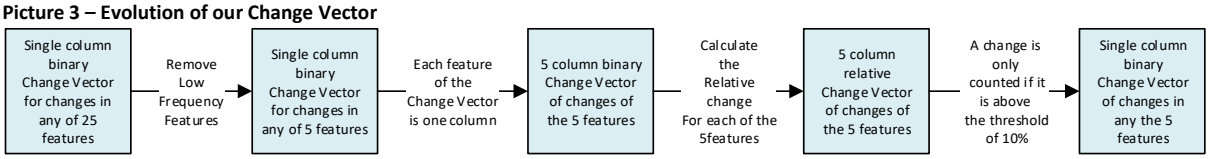
Training the RandomForest with this 5 column change vector enabled the prediction for each record and of the 5 individual predictor features if it requires a change of value.

The change frequencies for the test data 8Q2 are shown in Table 1 and for the Training Data 8Q3 in Table 2.

However, prediction results were increased by making a single column binary change vector out of the 5 columns where a change of any of the 5 features was counted as a change. The cost of this increased accuracy is that the individual suspicious feature is not identified.

This resulted in these numbers of labelled cases:
Training Data: 8Q3 has 3059 cases, 464 labelled as Change, reduced to 384 with 10% change threshold
Test Data: 8Q2 has 2912 cases, 474 labelled as Change, reduced to 387 with 10% change threshold

Picture 3 – Evolution of our Change Vector



In Picture 3 is our journey shown of finding the best Change Vector solution for this pilot study.

2.3 Feature Selection

The number of available training cases is rather small with 3059 in this pilot study and with the number of available features in a household record in excess of 2000, feature selection is an important part. Using all of them with the relatively small number of records would create too much noise for a machine learning algorithm.

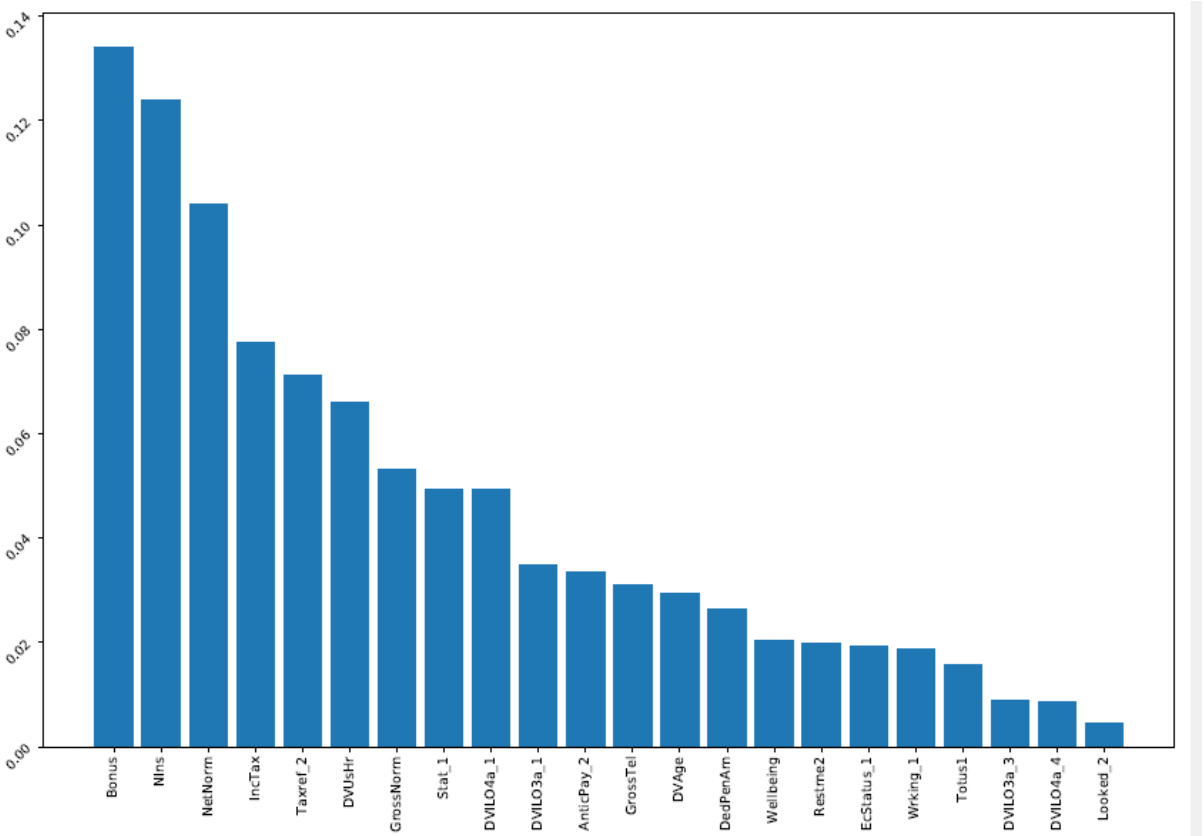
Initially, 91 features were selected from the topics listed in 2.2. These include 55 categorical features and with One-Hot-Encoding (OHE) the number of features increases to 246 features.

Table 3 – Feature Importance

Feature	Importance	Description
Bonus	0.134	Number of Bonuses in last 12 months
NIns	0.123	National Insurance paid over given period
NetNorm	0.104	Annual amount of net income (after deductions)
IncTax	0.077	Income Tax paid over given period
Taxref_2	0.071	Last pay did not include a tax refund
DVUsHr	0.066	Total number of hours worked over a week
GrossNorm	0.053	Annual amount of gross pay (before deductions)
Stat_1	0.049	Employee
DVILO4a_1	0.049	In employment
DVILO3a_1	0.034	In employment
AnticPay_2	0.033	Paid previously
GrossTel	0.031	Total income before deductions
DVAge	0.029	Age of person
DedPenAm	0.026	How much pension has been paid over given period
Wellbeing	0.020	Aggregate of four quality of life features
Restme2	0.019	How many years has the person resided at address
EcStatus_1	0.019	Working full time as an employee
Wrking_1	0.018	Paid work carried out in last7 days
Totus1	0.015	Hours per week usually worked
DVILO3a_3	0.008	Economically Inactive
DVILO4a_4	0.008	Economically Inactive
Looked_2	0.004	Has not looked for any paid work in last 4 weeks

Features with underscore followed by a number, e.g. Stat_1 is option 1 of the OHE Stat feature.

Picture 4 - Features



Experimenting with Feature Importance found that by using only features with an importance > 0.01, resulted in best model performance. The process used was this:

1. The RandomForest was first trained with all 246 features.
2. A list of 22 features with an importance > 0.01 was made.
3. The Random Forest was trained again, but this time only with the 22 features from the step above and their importance was calculated again. These 22 features are shown in Table 3 and Picture 4 with their new importance. The last 3 in Table 3 show an importance of < 0.01. This is their new importance from the RandomForest that was trained on only the 22 features selected previously

2.4 Output data

The prediction result for the 5 predictor features is shown in Table 4.

Table 4 – Visualisation of Prediction Result

Record ID	GrossNorm	NetNorm	IncTax	NIns	DedPenAm
2	12300	11650	40	40	
3	1892	1892	0		
5	6530	6530	0		22
20	12200	10616	28	55	
56	67656	44352	1198	305	
57		14897			
94		15634			
98		16730			
104	23450	23450			

Not real survey values

The prediction method of the Random Forest gives for each record a pair of numbers, e.g. [0.26, 0.64]. These are the "fake probabilities", or better described as voting scores of the 1000 trees used for that one record. In this case 260 trees voted for No-Change and 640 voted for Change.

With a prediction threshold of 35%, this record will then be counted as Change. The darker the colour in Table 4, the higher the voting score for that record to be above the threshold that it belongs to the Change class.

3. Machine Learning Solution

3.1 Models tried

Only ML algorithms from the sklearn Python library for supervised learning were used:

- DecisionTreeClassifier – this was used in early experiments to visualise a Decision Tree and extract rules the tree has built from the data pattern. But prediction results were poor and the extracted rules did not prove to be useful to the clerical editing team.
- Supervised NeuralNetwork – this did not perform well and was not perused any further.
- RandomForest – this proved to be very successful, please see below for results.

3.2 Model(s) finally selected and the criterion

The model used for this pilot study was RandomForest from the sklearn Python library.
The best Hyperparameters for the 8Q3 Training data were found with a grid search and are:

 # Initialise the Random Forest

```
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)
```

3.3 Hardware used

The hardware used to develop this pilot project was:

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

3.4 Runtime to train the model

Training of the model with 3056 cases and 22 features took 2.71 seconds

4. Results

For this pilot study the Change/No-Change classes are hugely imbalanced since the purpose is to detect only the cases where a data change was made by the clerical editing team.
With imbalanced datasets, it is possible for a model with even poor predictive capability to still have a high accuracy score by simply 'predicting' the negative class for each case, it is the more likely outcome and in the majority of cases it will be correct. For the Test Data the baseline Accuracy A is:

$$\mathrm{A} = \frac{TN}{TP + TN + FP + FN} = \frac{2912 - 387}{2912} = 86.7\%$$

and this is achieved even if all records are predicted to belong to the No-Change class. (2912 cases with 387 labelled as Change)

Consequently, in this type of scenario, recall and precision are more useful ways of assessing model performance.

Recall can be summarised as a measure of a model's capacity to correctly classify all the relevant cases (i.e., true positives) that exist within the dataset, while precision can be summarised as the model's capacity to accurately predict relevant cases, i.e., what proportion of its predictions were actually correct.

Each metric takes a different perspective on the algorithm's performance:

- Recall looks to minimise false negatives
- Precision looks to minimise false positives.

Consequently, there's a tension between Recall and Precision, and so a trade-off is needed.
Finding this balance between Recall and Precision is a judgment call that needs to be informed by the end-user's priorities. For this reason, it is often more helpful to frame questions about Recall and Precision in terms of user priorities.

Prioritising Recall could result in the teams not being able to save as much time and labour as hoped, because editors will have to spend some time reviewing irrelevant cases, while high Precision might result in a drop in data quality that might unsettle end-users of the survey data.

The tension between Recall and Precision makes it clear why the question 'what is accurate enough?' needs to be answered.

Table 5 – 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

The ML model predicts the score of a case falling in the Change class.
Only cases with a score above the Prediction Threshold are counted as belonging to that class.

Prediction results at the various Prediction Thresholds are shown in Table 5. The total number of predicted cases for each threshold is the sum of the True Positive (TP) and False Positive (FP) cases, TP + FP. The table shows how this sum decreases for increased thresholds.
The higher the threshold, the more certain the ML algorithm is that a case belongs to the Change class.

But as Table 5 also shows, with increasing thresholds, both the number for TP and the number for FP drop, but the numbers for FP drop faster. The F1-score, the harmonic mean of Precision and Recall and seen as a much more appropriate quality measure, peaks at around 50% threshold. This can be seen in Picture 5 as well, but at this threshold, Recall with a value of 74.4% might not be high enough.

The question: "What is good enough?" has not been answered yet.

It is not clear yet what the survey teams' accuracy thresholds are, or how the teams determine those thresholds. For example, what is the statistical significance of all the adjustments made as part of the Clerical editing processes compared to 'only' making those changes to the ML predicted records?

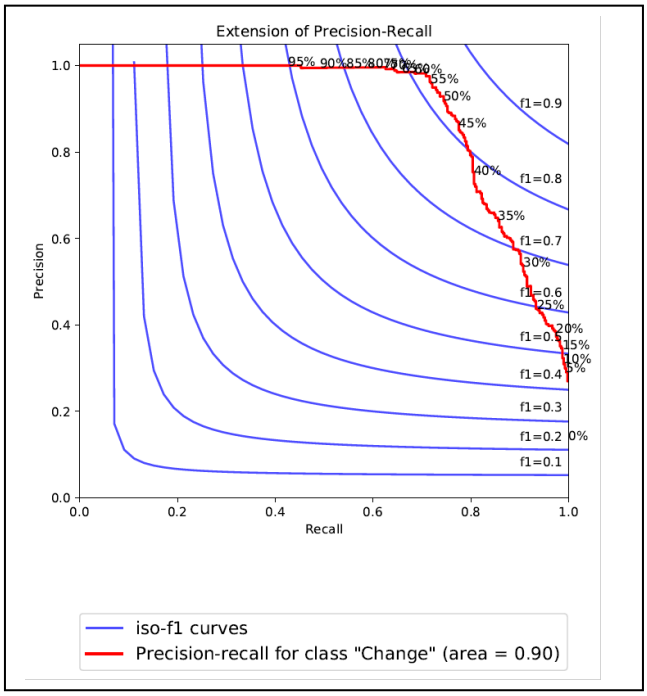
This baseline has not been established yet, but its understanding is obviously important for setting appropriate recall and precision thresholds.

For this pilot study the algorithm was trained on labels derived from records with a relative change of at least 10% in at least one of the 5 features, see Table 2. We need to find out if this type of selection and/or selective editing can be used to find the records whose changes in values will contribute the most to the quality of the data.

Discussions around this issue have started and are also seen to be of vital importance during the planning phase for operationalising this ML solution and for moving this pilot study towards implementation.

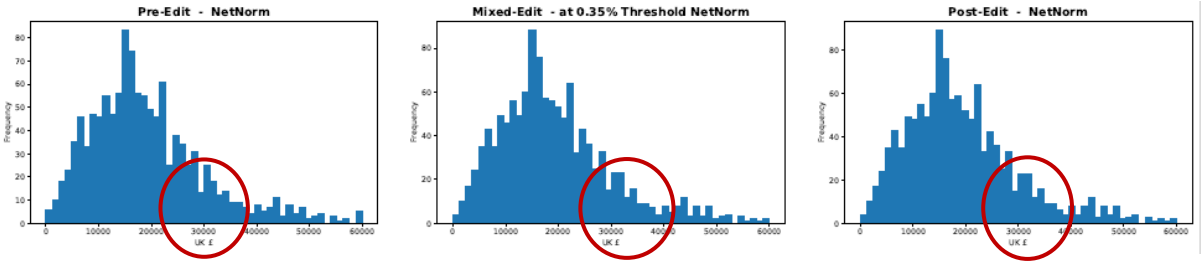
Also, given that there is a strong possibility the baselines will be different for the three surveys, there will need to be a discussion about the best way to derive harmonised thresholds.
The survey teams are not yet in a position to provide a steer about their business priorities. These are gaps that will need to be addressed when this pilot progresses to operationalisation.

Picture 5 – Precision-Recall Curve



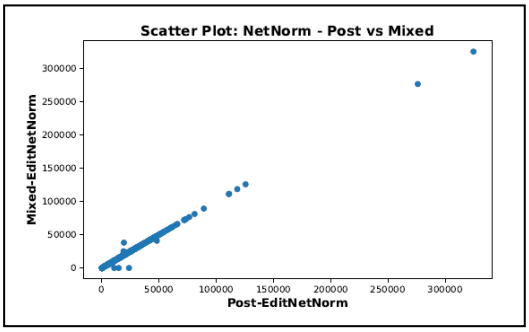
The Precision-Recall curve in Picture 5 shows how steeply the Precision decreases at a prediction threshold of about 55% with increasing Recall. There needs to be an informed discussion between the model builders and the survey teams to align recall and precision values with business priorities.

Picture 6 – Data Comparison at 35% Prediction Threshold

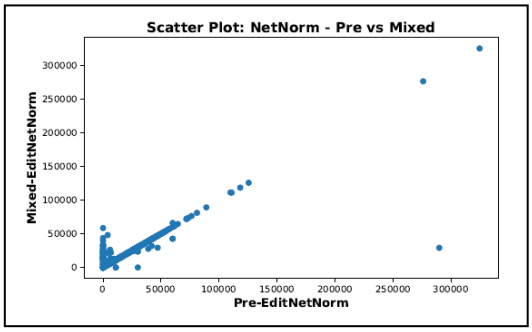


The 3 histograms in Picture 6 show the data distribution of NetNorm, the annualised value of NetPay, for the Pre Edit (raw data), the Post Edit (output from the clerical editing and imputation). The diagram in the middle shows the Mixed, a better name for this would be Simulated Output. The Simulated Output is a data set built, where the predicted records of the Pre Edited data set were replaced with Post Edited data. This the closest we can get to an ML output currently. The red circles show areas in the histograms where the Simulated Output is very similar to the Post Edited data but differs from the Pre-Edited data. This a good sign, but a statistical analysis has yet to be carried out.

Picture 7 – NetNorm Scatter Plot



Picture 8 – NetNorm Scatter Plot



The dots in the scatter plot in Picture 7 show where the Simulated data differ from the Post Edited data set. There appears to be only very few records that were not predicted but required a change of value for NetNorm and these are mostly very close to the straight line. This indicates that there are only small differences between the Post Edited and Simulated data sets.

But Picture 8 shows where the Simulated data differ from the Pre Edit data set. This shows that the Simulated data set has many more differences to the raw data and that these are larger in value as well.

Picture 6,7 and 8 show evidence that the Simulated data are very similar to the Post Edited data set and this suggests that the ML algorithm has predicted very well the records that need a change of NetPay and or NetPd values.

5. Code/programming language

This pilot study was programmed in Python in the JupyterNoteBook environment.
The code will be shared on the UN GitHub.

6. Evolution of this study inside the organisation

This pilot study was started out of curiosity, as an experiment to learn about ML and to find out if data records with data anomalies can be identified. Many years of experience in software development and support for the LCF survey made this survey an ideal candidate for this investigation.

This investigation was initially not driven to find a solution to a problem. However, as discussions with the survey teams progressed and early promising results indicated that this prediction is possible, the survey teams expressed their interest into this pilot study and their need to find a new way of identifying data records that need error correction for the emerging HFS, in fact, a new method has to be found for the HFS to go ahead.

This pilot study has so far only used LCF data as the raw and clerically edited data were readily available and accessible.

We collaborated extensively and learned from the LCF, SLC, WAS and HFS survey teams and their sponsors, the LCF clerical editing and imputation team in Titchfield, various methodology teams and the Editing and Imputation Expert Team at ONS.

A big step toward the implementation was the 2-day Strategic Engagement Workshop with Cloudera/Fast Forward labs on the operationalising of this ML pilot in March 2020.
This investigated:

- The legacy software and system infrastructure
- DAP (Data Access Platform), the future environment for all survey processing and production
- Transformation from legacy to DAP and where this pilot could fit in
- Short-term and medium-term roadmap for the implementation of this pilot

7. Is it a proof of concept or is it already used in production?

This pilot study is still a proof of concept, but the momentum is building up to work towards production. The results we already have and are shown above will help to compare accuracy/recall/precision from this pilot study with the legacy system once those figures have been established.

7.1 What is now doable which was not doable before?

This pilot study has shown that data records can be predicted to a high level of confidence for clerical error correction and thus reducing the need for the clerical team to look at data records that do need their attention. But all this relies on not carrying out less frequent and less significant changes.

If further analysis can show that this is statistically acceptable, a way forward out of the editing dilemma for HFS has been found.

7.2 Is there already a roadmap/service journey available how to implement this?

From the Cloudera workshop, we have identified these steps as a short-term roadmap that will aid the implementation:

- **Extend the predictive capability of the model:** The model's predictive capability needs to be extended to include anomaly detection for expenditure-related response fields. Alternatively, a separate model could be built to achieve the same purpose. This is a necessary step to avoid complicating the LCF team's workflow in ways that cancel out the time and labour related benefits from introducing ML.
- **Relax restrictions on development on legacy infrastructure**: As noted above, deploying the model will require some changes to the production environment. The immediate migration of the surveys' applications and data onto DAP would fit in with ONS' infrastructure transformation goals, but it is better to minimise the degree of change during this process. In order to implement this approach, the restrictions on (minor) development on legacy infrastructure will need to be relaxed.
- **Develop a plan for collecting baseline data for model evaluation purposes**: At present there are a number of gaps in the baseline data against which the model's performance (and by extension, value-add) can be assessed.
- **Establish thresholds for accuracy-related metrics:** As explained above, there's a need for an informed discussion between the model builders and the survey teams to align recall and precision values with business priorities.
- **Test model with WAS and SLC data:** To date the model performance has only been tested with LCF survey data, while there is an assumption that it will perform as well on SLC and WAS data, this has not been tested. This testing will become even more important if the predictive capability of the model is extended to cover expenditure-related anomalies as well.

Medium-term road map:

- **Develop protocol for implementing effective ongoing model monitoring and maintenance**: Initial deployment isn't the only consideration when it comes to integrating ML into the team's workflow. Model monitoring and periodic retraining are important for maintaining useable predictions from the model on an ongoing basis.
- Changes to survey questionnaires and tax/benefit policies have to trigger a model re-training. As well as the intervals for a model performance review has to be set, the performance threshold for model drift and the Responsibility for implementing these processes has to be clearly assigned.
  The ML technical team could be co-located with the survey teams. The UK government's approach to digital delivery provides a good precedent for the type of multidisciplinary teams.
- **Establish protocol for off- and on-lining the editing model when necessary**: There will be a need for periodical retraining. Clerical editing and imputation have to build a new training dataset and labelling and this will take some time. The ML solution has to be designed in a way that allow it to be switched on and off from the teams' workflows during this interval.

7.3 Who are the stakeholders?

The stakeholders implicated in our pilot study and its implementation are:

- The LCF, SLC, WAS and HFS survey teams
- SSOD – Social Survey operation Division
- HFS validation team
- LCF clerical editing and imputation team
- Data Science Campus
- Methodology
- Software and infrastructure architecture in DST – Digital Services and Technology
- All ONS internal and external data customers

7.4 Fall Back

A fall back mechanism has to be part of a production system based on this pilot study as stated above in section 7.2. New Training Data and from that new labels have to be created for the model to be retrained when policy changes to income/benefits are set by the government and/or the survey's questions change. The same fall back plan can then be invoked in case the ML solution fails altogether or if an unacceptable model drift has been detected.

7.5 Robustness

The question of 'How robust is the ML prediction in this pilot study' is difficult to answer. We have only been able to simulate the output of a ML driven process with the Simulated data set illustrated for NetNorm in Picture 6, 7 and 8. For this, all predicted records of the raw survey data for 8Q2 were replaced with data from the clerical editing process. A statistical analysis of that data set has not been carried.

Discussions with the LCF survey team has highlighted the issue that if this pilot study was to proceed towards implementation, it would have to interact with legacy systems and this would cause a problem with the current change freeze directive put in place during the transformation phase.

However, because the income question and data block for LCF and SLC are harmonised, it was decided to let this pilot study evolve into an implementation proposal for the SLC survey. We are now working on testing a ML model trained on LCF data to predict SLC records. Those predicted SLC records will then be put through clerical editing and will then be analysed by the survey team. Transfer learning, where a ML model is trained on a training data set and then used for prediction elsewhere will hopefully provide useful predictions.

In addition to this, the short-term and medium-term roadmaps as shown in 7.2 address the importance of robustness issues like model monitoring of the prediction quality.

8. Conclusions and lessons learned

ML can be used for editing, but we have to bear these points in mind:

- A new ground truth/gold standard data set for retraining the model has to be made periodically.
- The ML solution might not be available for the whole of the survey year, when from 1st April, the start of the survey and financial year, tax and benefit rules change and a new model has to be trained with new ground truth data. Survey data from the next 3 months will then have to go through the clerical editing process to form a new ground truth and during that period the ML model would not be available for those 3 months.
- ML expertise should be within the survey team to monitor and retrain the model when required.
- Editing will be far more efficient and faster with the ML solution compared to existing processes.
- Survey data will be available sooner for further processing and this will allow for more timely data and faster release.
- A cost savings analysis has not been done yet and we do not know if ML can save cost here, because clerical editing resources have to be maintained as well as technical expertise to build, analyse and keep the ML solution in operation.

9. Potential organisation risk if ML solution not implemented

Without this pilot study being developed into production, the HFS survey has to either rely on the in-depth clerical editing approach currently applied to the LCF survey or on the scripted outlier detection system used for SLC and WAS.

An editing dilemma has been identified as for each of these approaches have their intrinsic shortfalls of over-editing or accuracy. For the HFS to proceed, a new editing solution that can work fast, efficient, consistent and reliable has to be found. And that can be achieved with ML.

Like any other NSI, ONS will get under more and more pressure to provide trusted statistical out in a fast evolving world where commercial competition is fast growing. To maintain it's relevance and trust, ONS has to embrace further new technologies like ML.

10. Has there been collaboration with other NSIs, universities, etc?

Continued collaboration with DeStatis (Germany) and Istat (Italy) has helped to form this pilot studies.
And the early in-house collaboration with Cloudera has helped to build the early proof of concept.

11. Next Steps

To drive this project onwards with senior engagement, a number of meetings have been setup on senior level with the Architecture and the social survey teams.

A short term solution is required for the HFS to proceed with the harmonisation of income data. A plan to replace the existing scripted outlier detection for SLC data with a ML solution is now under way. This will have no impact on other existing software or infrastructure implementations.

In addition, detailed discussions about the scope and timetables of the software uplift of all social survey systems, which are classed as legacy under a change lockdown, is under way. This will be key to decide where the ML solution can sit and what technology will be used.

The challenges highlighted in the road maps (see section 7.2) will be discussed soon. Adding expenditure data to the ML prediction will change recall and precision and experiments on this need to be carried out.