Edit and Imputation With Machine Learning from PoC on LCF towards Implementation for SLC/HFS survey data

Claus Sthamer
Data Science Campus

2nd September 2020
### International Definitions:

**Edit:**  
Identifying records that need values changed or missing values inserted

**Imputation:**  
Changing and Inserting missing values

### ONS Social Survey Definitions:

**Edit:**  
Correct an incorrect value

**Imputation:**  
Inserting missing values
Pre-Processing & Testing

Change Vector
25 Features from LCF Editing Instruction Manual

Keep removing features with low change frequencies → 5

Hyperparameters

Grid search to find best parameters

Train

Better Understanding of categories

Predict

One Hot Encoding

Evaluate

Good Prediction at higher Threshold leads to higher confidence

106 Data Features:
Education
Income
Family composition
Job and secondary Job
Benefits

Keep removing features with low Importance → 18

Prediction Threshold

Dendograms

Figure out strongly correlated features

Eliminate small changes

Change Threshold = 10%

Keep removing features with low change frequencies → 5
Results

Change vector with number of changes:

- NetPay: 81
- IncTax: 235
- NLns: 254
- GrossPay: 336
- DedPenAm: 113

Training Data: 8Q3 3059 Person records, 442 labelled as Change, reduced to 362 with 10% change threshold
Test Data: 8Q2 2912 Person records, 451 labelled as Change, reduced to 361 with 10% change threshold

Features:
106 features reduced to 18 by removing features with low Importance

Prediction Threshold: 20% 25% 30% 35% 40% 45% 50%
Recall 97.2% 94.6% 88.1% 84.0% 78.6% 75.2% 72.9%
Precision 35.7% 40.0% 45.4% 54.5% 67.1% 81.3% 90.1%
F1-Score 52.2% 56.2% 59.1% 66.1% 72.4% 78.1% 80.6%
TP 376 366 341 325 304 291 282
FP 677 550 410 271 149 67 31
The Journey so far:

UNECE ML Project
To improve:
- Time lines
- Accuracy
- Cost

Editing of LCF with ML POC

- Faster
- 75% Savings
The HFS and it’s component Surveys

LCF
Expenditure & Nutrition
5000

SLC
Living Conditions
12000

WAS
Wealth & Assets
10000

HFS
27000

Number of co-operating Households

Claus.Sthamer@ONS.gov.uk
Survey Specific Editing (Currently in Production)

LCF – all cases go through clerical editing
too slow and too labour intensive

Speed & Cost? Over Editing?

SLC – Scripted outlier detection (range of values)
Only about 10% of changes that are made with the LCF method are made with the SLC method

Accuracy?

ML

WAS – Scripted outlier detection (range of values)

Accuracy?
Focus has shifted to SLC survey data

- Income Statistics for LCF & SLC differ
- Possible reasons:
  - Different Edit processes?
  - Different Imputation?
  - SLC is Longitudinal – older respondence tend to stick with it

- Which Editing process should be adapted for all 3 surveys?
What are the Challenges?

1. Business Knowledge
   - Stakeholders - Who can make decisions?
   - Data usage - Impact on data users
   - Existing Data Pipeline - Integration

2. Knowing the Data
   - Precision ↔ Recall: to satisfy business needs

3. Value Proposition, what is the ML value?
   - Speed?
   - Cost?
   - Accuracy? What is the driver for all this?
Strategic Engagement Outline

● ‘deep dive’ discussions about infrastructure, survey pipelines, etc.

● apply an evaluation framework to a ML PoC designed to understand both its strengths and the challenges it faces in progressing towards operationalisation.

● recommendations for addressing these challenges in the form of short and medium-term roadmaps.
Short term roadmap

1. Extend the predictive capability of the model.

2. Relax restrictions on development on legacy infrastructure

3. Develop a plan for collecting baseline data for model evaluation purposes: high workload?

4. Establish thresholds for accuracy-related metrics

5. Test model with WAS and SLC datasets
Medium term roadmap

1. Develop protocol for implementing effective ongoing model monitoring and maintenance.

2. Establish protocol for off- and on-lining the editing model when necessary.

3. Incorporate an interpretability framework and explainability approach.
End-to-end Machine Learning in an Enterprise setting

Production ML is an iterative team sport

[Diagram showing the process of ML in an enterprise setting]

- Business Inputs
- Data Science
- Model Training
- Packaging
- Deploy and Serve
- Monitor

- Business Validation
- Data, Metadata, Security, Workload Management, Governance
Other the Challenges?

• Technology
  • DAP
  • IDP

• Google, AWS, Azure

• Who will do what?

• Ethics & Governance
The Journey so far:

UNECE ML Project
To improve:
• Time lines
• Accuracy
• Cost

Editing of LCF with ML
POC

Faster
✔️ 75% Savings

CFFL
Strategic Engagement

Editing of SLC with ML

1st SLC Experiment

Office for National Statistics
Claus.Sthamer@ONS.gov.uk
Thank you