Data Science Campus

Machine Learning for Social Surveys LCF/SLC/HFS

Claus Sthamer
Technical Project Lead

26th January 2022
What is Machine Learning?

Machine Learning is used to identify rules and patterns in data humans can not.
**Example: How to recognise a cat**

But not just one cat but all cats. What are the rules?

A ML algorithm can learn from pictures as long we tell it (Labels) what they are.

ML can make an inference of the class of new pictures, it gives a score for the most likely class.
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 each year
Survey Specific Editing (Currently in Production)

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

- **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 50% (?) of changes that are made with the LCF method are made with the SLC method
  - Accuracy?

- **WAS** – Scripted outlier detection (range of values)
  - Accuracy?

Data Science Campus
Editing of LCF income data with ML

LCF Editing Instructions for Income → extensive manual Editing → Ground Truth

5 most often changed Independent Variables:
- Net income (after deductions)
- Income Tax
- Gross pay (before deductions)
- National Insurance paid
- Deduction for pension

Any change of more than 10% is counted as a change → Label/Target/Dependant

Two Class Classification Problem:
- No-Change
- Change

Training data → Train the algorithm
Test data → Test the trained model

Supervised Learning

Random Forest ML algorithm
Results of 2912 LCF Test Cases – What is good enough?

Battle between Recall and Precision, they can’t both be 100%

<table>
<thead>
<tr>
<th>Prediction Threshold</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
<th>40%</th>
<th>45%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>95.7%</td>
<td>94.0%</td>
<td>91.8%</td>
<td>88.9%</td>
<td>84.2%</td>
<td>81.0%</td>
<td>77.4%</td>
</tr>
<tr>
<td>Precision</td>
<td>37.0%</td>
<td>41.3%</td>
<td>47.5%</td>
<td>55.1%</td>
<td>61.8%</td>
<td>69.8%</td>
<td>77.4%</td>
</tr>
<tr>
<td>F1-Score</td>
<td>53.3%</td>
<td>57.4%</td>
<td>62.6%</td>
<td>68.0%</td>
<td>71.3%</td>
<td>75.0%</td>
<td>77.4%</td>
</tr>
<tr>
<td>TP</td>
<td>352</td>
<td>346</td>
<td>338</td>
<td>327</td>
<td>310</td>
<td>298</td>
<td>285</td>
</tr>
<tr>
<td>FP</td>
<td>600</td>
<td>491</td>
<td>374</td>
<td>267</td>
<td>192</td>
<td>129</td>
<td>83</td>
</tr>
</tbody>
</table>

\[
\frac{327}{368} = 88.9\%
\]

If prediction score > Threshold \(\rightarrow\) Case belongs to the Change Class

Objective:
1. Find as many True Positive (TP) as possible with small number of False Positives (FP) - 88.9% ✓
2. Find all cases with large value changes > 500% - 100% ✓
3. Reduce the number of cases to be manually analysed (from 3000 to about 600) - 80% ✓

The same team could check 5 x the number of cases for inconsistent data \(\rightarrow\) HFS

Data Science Campus
Then we looked at SLC survey data

• SLC also collects Income data, but uses a deterministic editing process

• No SLC Ground Truth, but we know which cases have been changed (Labels)

• Transfer Learning – Can we use the trained LCF ML model?

• Objective:
  • Reduce: number of cases flagged by the deterministic SLC process, but do not receive a value change
  • Keep all cases with Changes
  • Keep all cases with large value changes
Proposed SLC Editing Pipeline

Objectives:

- Retain the SLC Editing method
- Add ML as a filter to remove cases that: Are flagged by the scripts, but do not need any changes → ‘Wasted Effort’
- Reducing this Burden is key to use the SLC Editing method for HFS
- Retain cases that
  - need changes
  - need large value changes
- A longer term objective is to have only one Editing process
### Results of 6084 SLC Cases – What is good enough?

**Battle between Recall and Precision, they can’t both be 100%**

#### SLC Prediction Results:

<table>
<thead>
<tr>
<th>Prediction Threshold</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
<th>35%</th>
<th>40%</th>
<th>45%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recall</strong></td>
<td>99.7%</td>
<td>99.7%</td>
<td>99.3%</td>
<td>98.4%</td>
<td>97.4%</td>
<td>95.2%</td>
<td>92.6%</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>34.5%</td>
<td>35.8%</td>
<td>38.7%</td>
<td>43.7%</td>
<td>52.2%</td>
<td>60.9%</td>
<td>69.4%</td>
</tr>
<tr>
<td><strong>F1-Score</strong></td>
<td>51.2%</td>
<td>52.7%</td>
<td>55.7%</td>
<td>60.6%</td>
<td>67.9%</td>
<td>74.4%</td>
<td>79.3%</td>
</tr>
<tr>
<td><strong>TP</strong></td>
<td>753</td>
<td>753</td>
<td>750</td>
<td>743</td>
<td>734</td>
<td>719</td>
<td>699</td>
</tr>
<tr>
<td><strong>FP</strong></td>
<td>1432</td>
<td>1351</td>
<td>1188</td>
<td>956</td>
<td>674</td>
<td>462</td>
<td>308</td>
</tr>
<tr>
<td><strong>Burden cases filtered out</strong></td>
<td><strong>76%</strong></td>
<td><strong>82.6%</strong></td>
<td><strong>85.3%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

736 of 968 | 800 of 968 | 826 of 968

If prediction score > Threshold → Case belongs to the Change Class

#### Objective:

1. Find as many True Positive (TP) as possible with small number of False Positives (FP) - 95.2%

2. Find all cases with large value changes >300% - 100%

3. Reduce the number of cases manually checked, without receiving value changes - 82.6%

(240 hours annually)

The same team could check 5 x the number of cases for inconsistent data → HFS
Challenges

Results show that ML can make data Editing more efficient

Questions:
• What is good enough?
• What is it we want to achieve?
• How do we put this into production?
• How will ML fit into the data pipeline?

What about:
• Training Data
• Model/Data Drift
• Ethics?
• Explainability?
The Research and Implementation ML Workshop Part 1 will look at these in more detail

Technical ML sessions will look in more details how ML works and how to do it.

Thank you