The imputation of the “Attained level of Education” in the base register of individuals: an experimentation using Machine Learning techniques

Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto
**HMG-MOS Machine Learning Italy pilot**

**THE AIM**
Determine how and where Machine Learning techniques (ML) can give greater benefits in solving the imputation problems compared with classic statistical models.

**Target Variable and Data**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage</td>
<td></td>
<td></td>
<td>A (8%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B (89%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C (4%)</td>
<td></td>
</tr>
</tbody>
</table>

About 5% with Target variable
Different imputation steps (due to the complexity of available information and different patterns).

A: $P(\text{ALE18} \mid \text{ALE17}, \text{age18}, \text{citiz18})$
B: $P(\text{ALE18} \mid \text{ALE17}, \text{age18}, \text{citiz18}, \text{prov18}, \text{gender})$
C: $P(\text{ALE18} \mid \text{age18}, \text{gender}, \text{citiz18}, \text{apr})$
METHODS:

ML technique: Multi Layer Perceptron (MLP)

- All available variables
- One imputation step
- Dummy representation
- No pre-treatment

- two hidden layer with 128 neurons
- fully connected
- dropout
- deep learning framework

KERAS
RESULTS:

Comparison between target and estimated distributions

Legend:
- target
- estim.
- target=estim.

Log-lin.

MLP
RESULTS:

Estimated ALE distributions for individuals with a PhD (item 8)

Log-lin.

MLP

Bar length \propto \text{frequency of predicted ALE}
RESULTS:

Micro-level accuracy: Log-linear vs MLP

<table>
<thead>
<tr>
<th>Fold</th>
<th>Target=estimated</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Log-lin.</td>
<td>MLP</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.722</td>
<td>0.735</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.721</td>
<td>0.736</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.723</td>
<td>0.737</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.721</td>
<td>0.735</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.721</td>
<td>0.734</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td>0.721</td>
<td>0.735</td>
</tr>
</tbody>
</table>

Model accuracy is calculated using the 5-fold approach.

Micro level accuracy of imputed ALE 2018 using ML technique is very similar to those originated from Log-Linear models: 73.5% vs 72.1%

variance of results is in both cases negligible.
CONCLUSIONS:

- The results of estimation with the two approaches are completely comparable.

- For particular sub-population, such as extreme items (PhD), Log-linear imputation is better.

- MLP micro accuracy is a bit better respect the loglinear model

- MLP approach does not require variables pre-treatment
Fabrizio De Fausti  e-mail defausti@istat.it