The imputation of the "Attained Level of Education" in the base register of individuals: an experimentation using Machine Learning techniques

Prepared by De Fausti F., Di Zio M., Filippini R., Toti S., Zardetto D., Istat, Italy

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: Attained Level of Education (ALE) for Italian resident population in 2018.

Data: The Base Register of Individuals (BRI) results from the integration of different sources (survey and administrative). A high amount of information on ALE is available, however, delay of information and coverage problems arise, hence micro data imputation/prediction is necessary.

Methods

Classic statistical model: Log-linear

Different imputation steps (due to the complexity of available information and different patterns).

A: $P(ALE_{18}| ALE_{17}, age_{18}, citiz_{18})$

B: $P(ALE_{18}| ALE_{17}, age_{18}, citiz_{18}, prov_{18}, gender)$

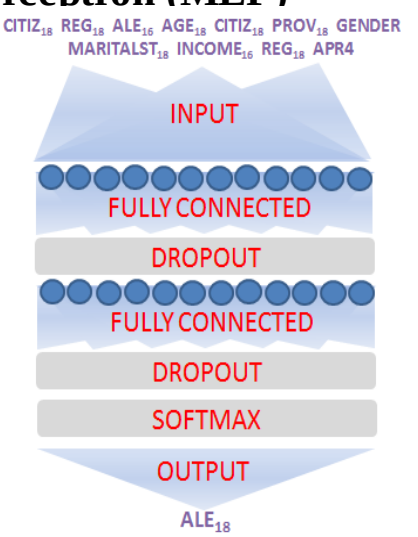
C: $P(ALE_{18}| age_{18}, gender, citiz_{18}, apr)$

ML technique: Multi Layer Perceptron (MLP)

One imputation step:

ALE_{17}, age_{18}, $citiz_{18}$, $prov_{18}$, gender, apr (no pre-treatment)

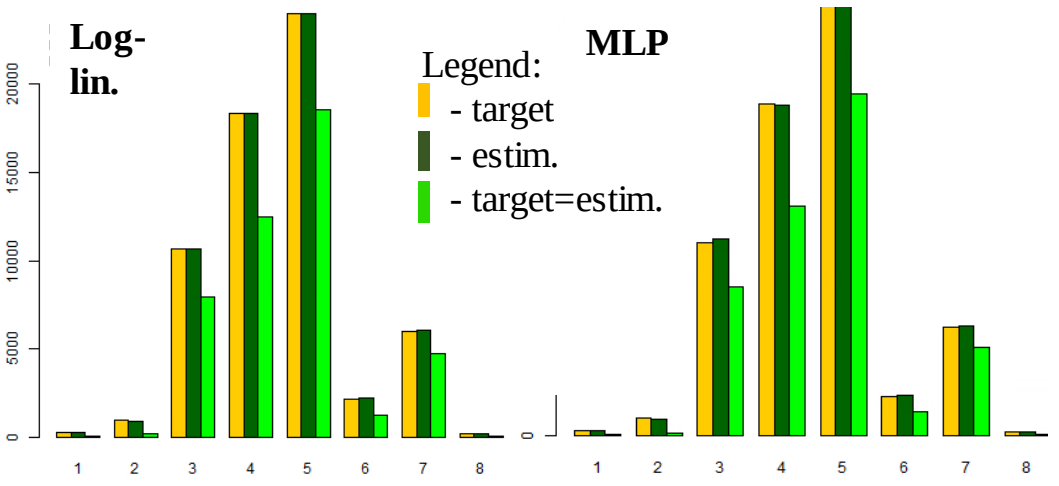
- Network composed by 2 hidden layer with 128 neurons fully connected.



The imputed value is a random value extracted from the probability distribution of correspondent pattern

Results

Graph 1: Comparison between target and estimated distributions



Graph 2: Estimated ALE distributions for individuals with a PhD (item 8)

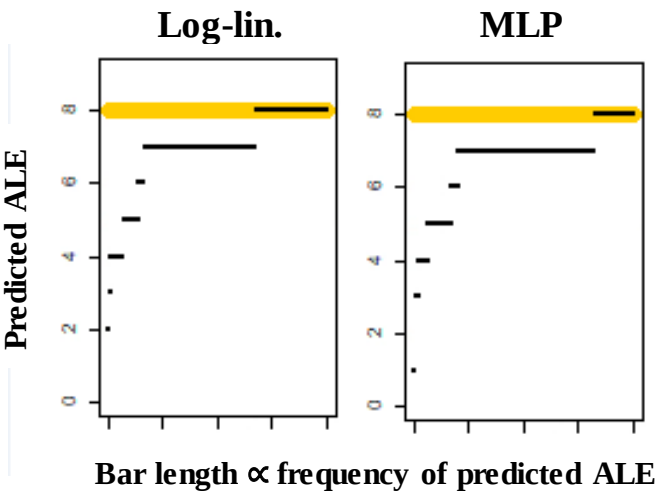


Table 1: Micro-level accuracy: Log-linear vs MLP

Fold	Target=estimated	
	Log-lin.	MLP
1	0,722	0,735
2	0,721	0,736
3	0,723	0,737
4	0,721	0,735
5	0,721	0,734
mean	**0,721**	**0,735**

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 aproaches are completly comparable.
- For particular sub-population, such as extreme items (PhD – graph 2), Log-linear imputation is better.
- MLP approach dos not requires variable pre-treatment