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

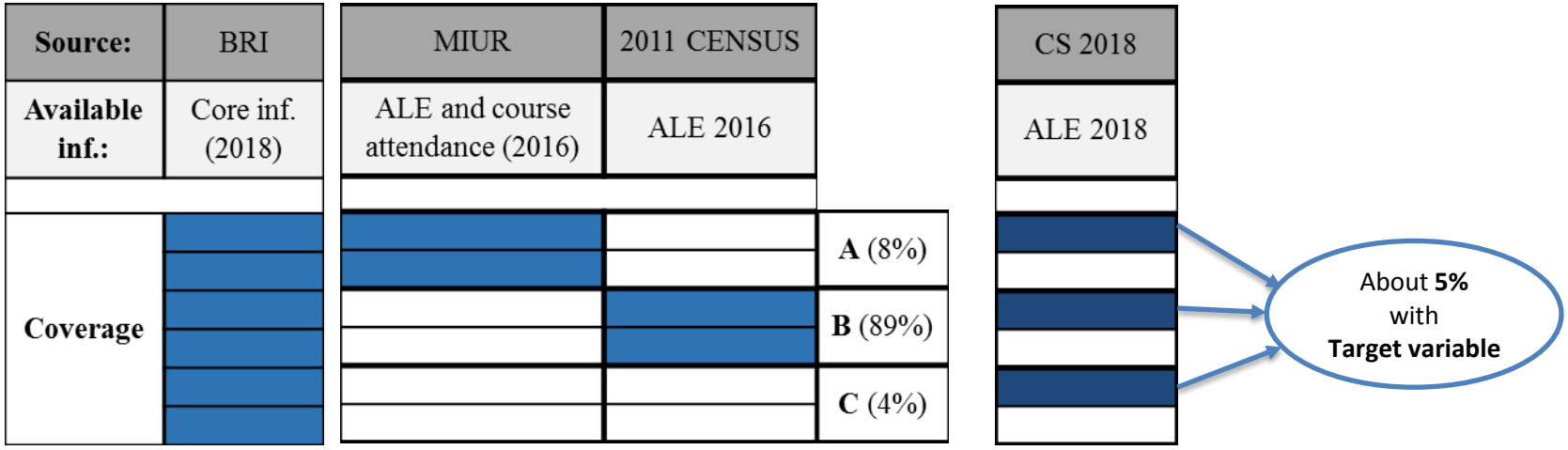
Fabrizio De Fausti, Marco Di Zio, Romina Filippini, Simona Toti, Diego Zardetto

HLG-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



METHODS:

Classic statistical model: Log-linear

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

A: P(ALE18| ALE17, age18, citiz18)

B: P(ALE18| ALE17, age18, citiz18, prov18, gender)

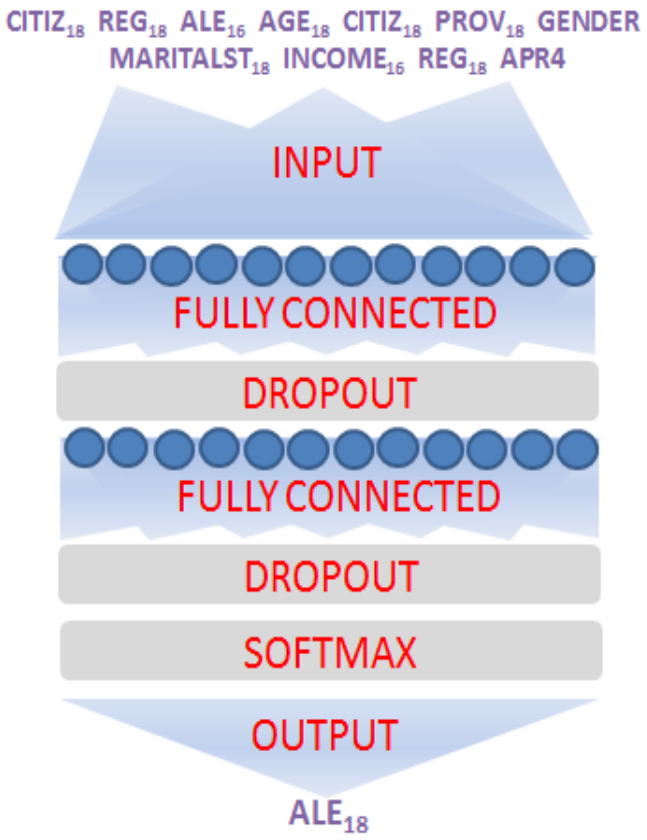
C: P(ALE18| age18, gender, citiz18, apr)

METHODS:

ML technique: Multi Layer Perceptron (MLP)

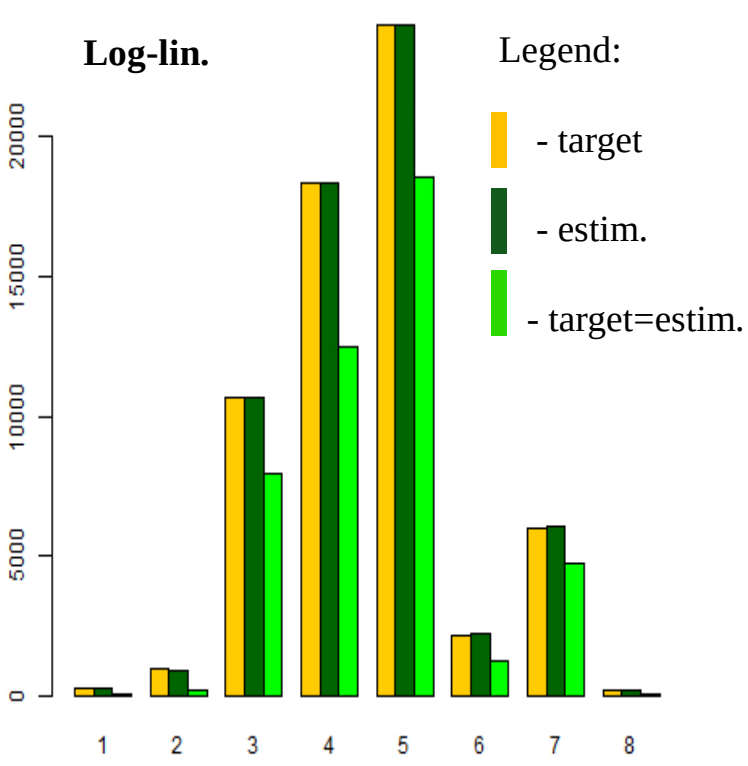
- **All available variables**
- **One imputation step**
- **Dummy representation**
- **No pre-treatment**

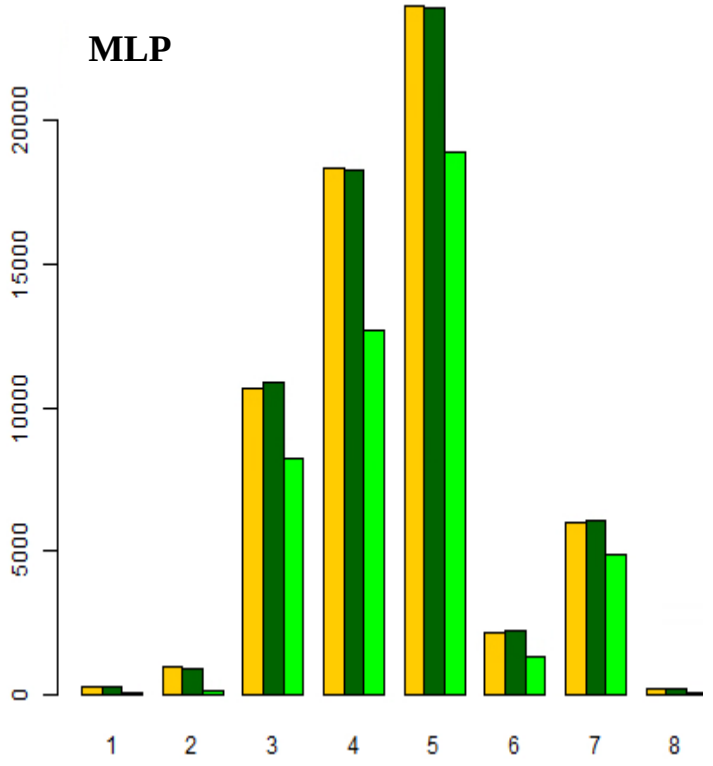
-two hidden layer with 128 neurons
-fully connected
-dropout
-deep learning framework KERAS



RESULTS:

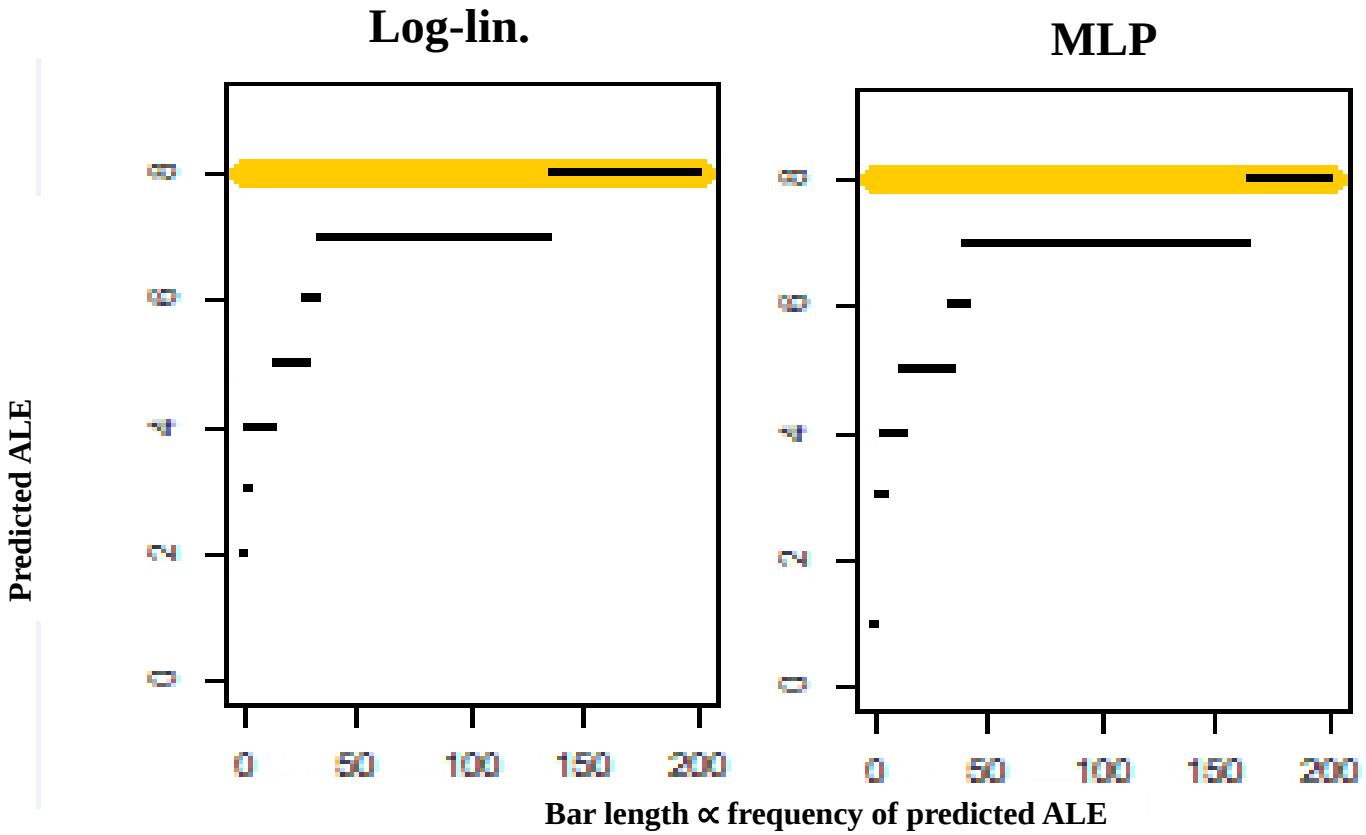
Comparison between target and estimated distributions





RESULTS:

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



RESULTS:

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), 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