UNECE HLG-MOS

Machine Learning for Official Statistics

Imputation of the variable "Attained Level of Education" in Base Register of Individuals

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

THE AIM

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

The Use Case

Type and source of data:

Data of different nature are jointly used:

- administrative data,
- traditional Census data
- sample survey data.

Source:	BRI	MIUR	2011 Census	CS 2018		Subsets selected to conduct the study
Available inf.:	Core inf.	ALE 2017	ALE 2017	ALE 2018	Sub-pop.	
Coverage					A	Yes
					A	No
					B	Yes
					B	No
					C	Yes
					C	No

Type and source of data:

Data of different nature are jointly used:

- administrative data,
- traditional Census data
- sample survey data.

Source:	BRI	MIUR	2011 Census	CS 2018		Subsets selected to conduct the study
Available inf.:	Core inf.	ALE 2017	ALE 2017	ALE 2018	Sub-pop.	
Coverage				OK	A	Yes
					A	No
				OK	B	Yes
					B	No
				OK	C	Yes
					C	No

Only one Italian region: Lombardia

The dataset for the experimentation consists of **312.813 individuals** with no missing data on **ALE 2018 (target variable)**.

Classic statistical model: Log-linear

For each subpopulation (A, B and C), the best Log-linear model is chosen so we obtain many models.

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

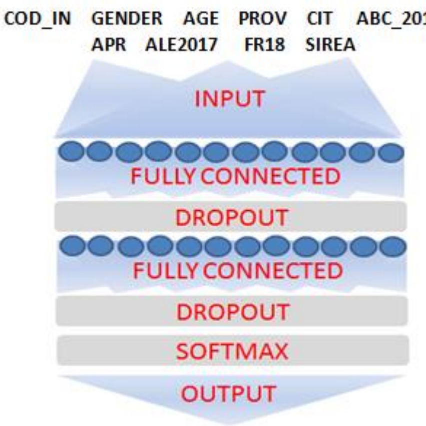
B: P(ALE18| ALE17, age18, citiz18, prov18, gender)

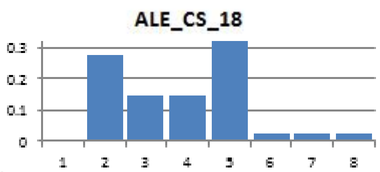
C: P(ALE18| age18, gender, citiz18, apr)

Methods

ML technique: Multi Layer Perceptron (MLP)

- Experience with NN for NLP and Image Recognition.
- Simple **neural network** architecture, the Multi Layer Perceptron (MLP), to find the approximation of the **relationship** between the **input** variables and the probability distribution of the **output** variable for each pattern.
- We **impute** the ALE item **randomly extracted** from the **probability distribution** of the correspondent pattern.



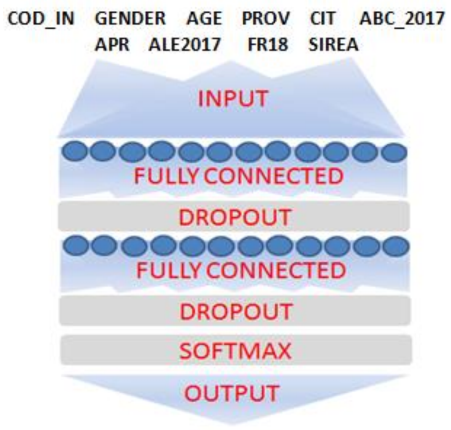


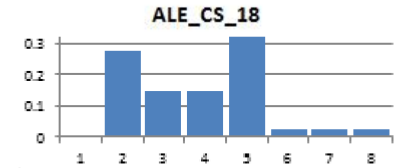
Model Training

- Dataset (312.813 samples) **splitting:** 80% Train and 20% Test
- **Input variables** are the **same** of LogLinear model
- Model selection: Best loss on Validation (20% of Train)

ML technique: Multi Layer Perceptron (MLP)

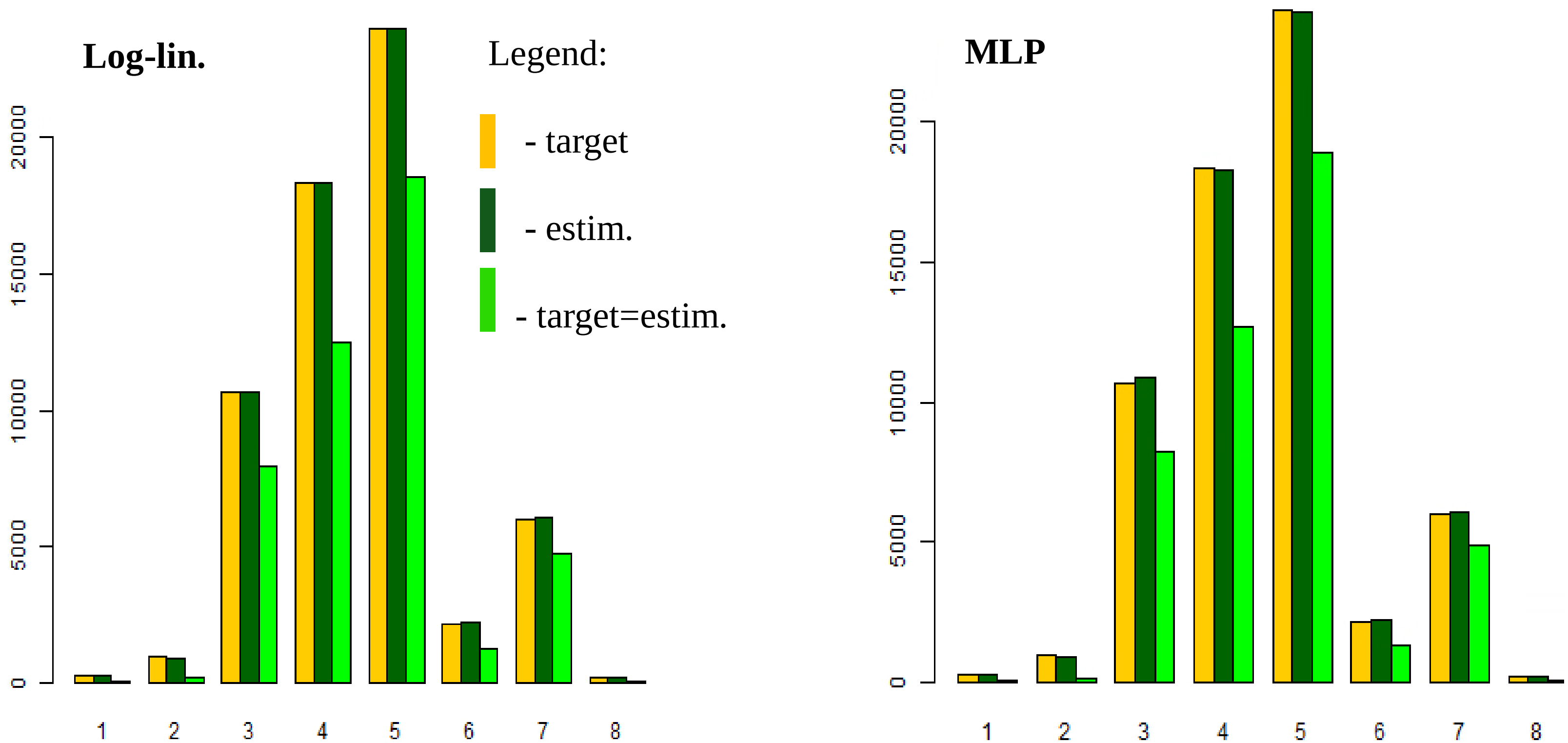
- Two hidden layer fully connected
- 128 neurons for each layer
- Dropout
- Softmax output layer
- Deep Learning framework KERAS





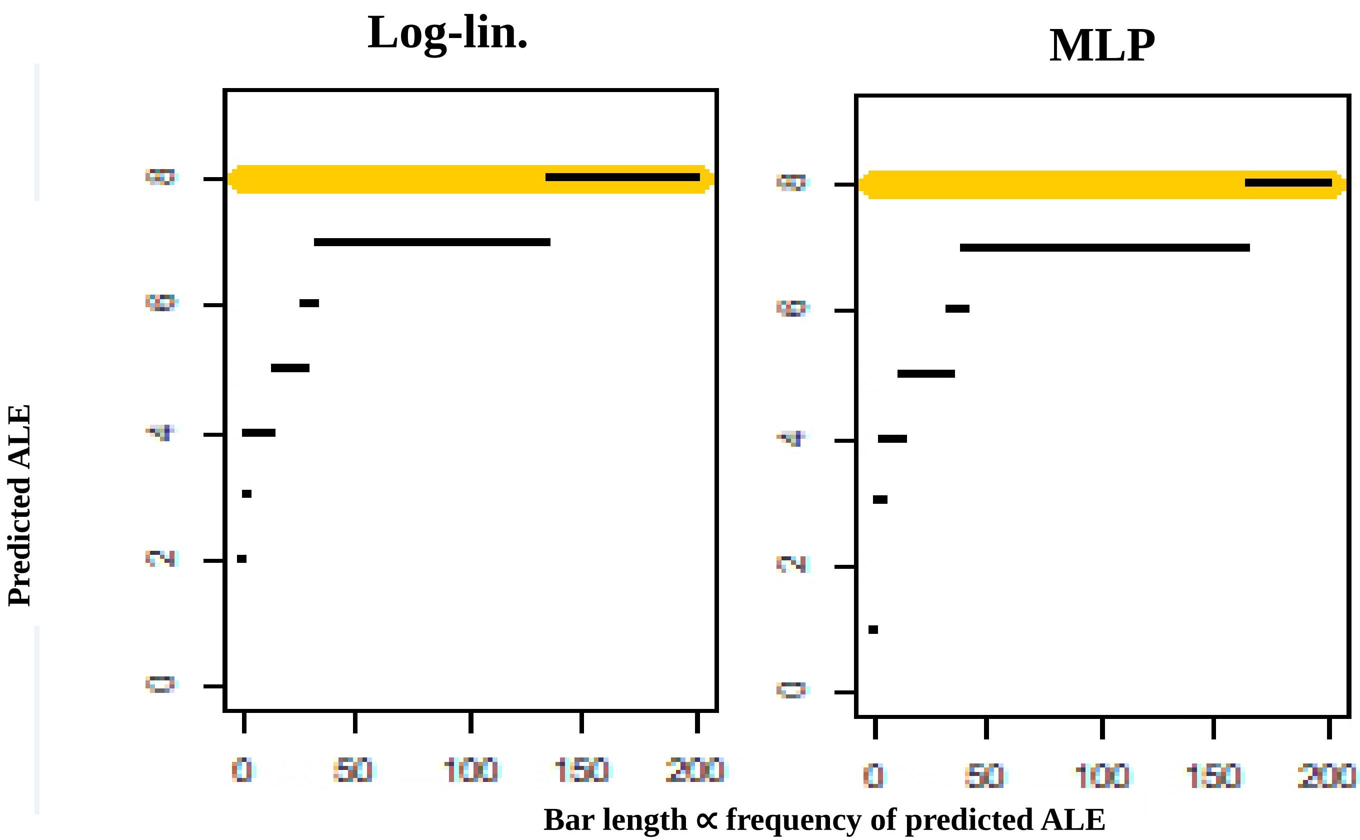
- All available variables
- One imputation step
- Dummy representation
- No pre-treatment

Comparison between target and estimated distributions



Results

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



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