3P-U – Debiaising using MICE

Émilien ARNAUD[1,2], Daniel Aiham GHAZALI[1], and Gilles DEQUEN[2]

[1]Department of Emergency Medicine, Amiens University Hospital
[2]Modelisation, Information and Systems laboratory, Amiens Picardy Jules Vernes University

Wednesday 23rd March, 2022

1 Rational

The hypothesis is that real measured values contain different biases : measure bias, human reporting biais, ...

MICE algorithm imputes missing values based on known values in the dataset. We hypothesize that the generated MICE model used to fill values contains no bias. Then, applying the model on the measured values would improve the model.

2 Setup

The setup can be run using this command : `cli.py setup`

2.1 Split indexes

We take the raw dataset, we random 20% indexes : the validation set. The rest is the train set. The validation will only be used to validate our experiments.

The `archived/validation_indexes.pkl` files stores the validations indexes. The train indexes are rest.

2.2 Missing data

For both dataset, we store all locations (index, column) were a data is missing

2.3 Columns description

The experimenter must define in `archived/columns_description.py` two variables to correctly type the dataframe columns:

1. `cols_categorical`: List of columns that should be considered as categorical
2. `cols_numerical`: List of columns that should be considered as numerical

3 Protocole

3.1 First setup

We begin to launch the setup command

3.2 Frist train of the raw data with no modification

We train a reference model on the train data. This is the base model. All other models will be trained using the same method to be compared. The generated model is presented **raw** in Figure 1

1. We load split dataset
2. We train a model on the training dataset
3. We compute the AUC on the validation dataset

3.3 Training MICE

1. We train MICE on training dataset only.
2. We replace missing values in both dataset
3. We generated a model which is presented **mice** in Figure 1

3.4 Replace non missing data

Here, we replace the non missing data by imputed data. We exclude from this treatment the target column and columns were all data were filled, since the completion model will not work.

Excluded variables from being debiased are : CIMU, AGE, SEXE, HOSPITALISATION, HEURE_ARRIVEE, JOUR_SEMAINE, MOIS, SEMAINE_ANNEE, main complaint one hot encoded

3.4.1 Delta between datasets

We compute the delta of the values $\Delta = V_{mesured} - V_{imputed}$. The differences are presented in Table 1

Question : how can we measure the delta for categorical variables ?

Variable	Min	Mean	Std	Max
CETONEMIE	-16	-0.0033	0.37	127
DOULEUR	-10	-0.09	3.41	10
FC	-170	0.09	21.72	188
GLYCEMIE	-102	-0.27	3.28	496
HEMOCUE	-12.98	-0.00	0.56	13.55
OH	-8.32	-0.00	0.16	7.62
OXYGENE	-126	-0.04	3.38	126
PAD	-935	0.24	28.21	928
PAS	-217	0.28	27.14	200
SATURATION	-45	0.02	2.60	44
TEMPERATURE	-37.6	-0.00	0.90	9.20

Table 1: Difference between the dataset filled by MICE and the dataset filled and corrected by MICE.

3.4.2 Model

The generated model is presented **mice_corrected** in Figure 1

3.5 Using deep learning

We trained a models using deep learning rather than XGBoost to see if the type of evaluation model is dependent to the MICE and correction.

3.5.1 For the raw dataset

To use deep learning method, we must fill all missing values, then we used mean for numerical variables and mode for categorical variables. The generated model is presented **deep_raw** in Figure 1

3.5.2 For the MICE dataset

We trained a deep model with no modification of the MICED dataset. The generated model is presented **deep_mice** in Figure 1

3.5.3 For the MICE corrected dataset

We trained a deep model with no modification of the MICE corrected dataset. The generated model is presented **deep_mice_corrected** in Figure 1

3.6 Comparing all models

We put all AUC on the same chart in Figure 1

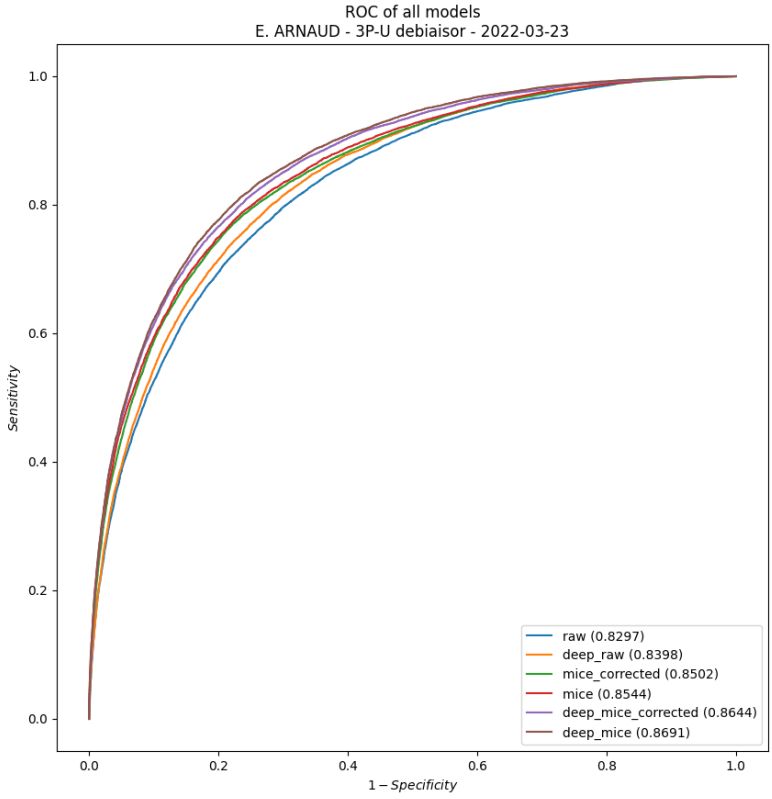

Figure 1: AUROC of the models in differents stage of the pipe line. All models starting with `deep_` refer to a deep learning model, others refer to an XGBoost model.