Data and Information Quality

Classification of imputed data

Authors:
La Greca Michele Carlo (10864460)

Professor:
Prof. C. Cappiello

February 10, 2023

Contents

1 Introduction **2**

2 Data **2**

3 Setup choices **3**
3.1 Imputation techniques . . . 3
3.1.1 Standard imputation . . . 3
3.1.2 Advanced imputation . . . 4
3.2 Imputation accuracy assessment . . . 4
3.3 Machine learning algorithms . . . 4
3.4 Machine learning evaluation metrics . . . 4

4 Results **5**
4.1 Imputation accuracy . . . 5
4.2 Machine learning algorithms performance . . . 7

5 Conclusion **7**

1 Introduction

Missing values are an important element to deal with when it comes to work with Machine Learning and in general with data. The presence of missing data may become a problem since the information are likely to be biased on the smaller amount of data. Moreover, ML methods suffer missing value since they don't accept null value, and also because they work well when a lot of data are available. The project aims to build a case study on the comparison between different types of imputation techniques, as well as measuring the performances of two ML algorithms for classification trained and evaluated on a dataset that each time has a bigger rate of missing values imputed.

First, different version of the dataset will be available, each with a different amount of missing values. For each dataset with missing values two types of imputations will be applied, and the resulting datasets will be compared with the original one with complete data. Finally, two machine learning algorithms will be applied to the complete dataset and to the imputed ones, measuring and comparing the performances obtained.

2 Data

The dataset is composed of 3017 records, each having 15 features, 10 are categorical and 5 are numerical. One of the categorical is intended to be the target that will be predicted by the algorithm. One variable has been removed since every row consisted of a unique value. Thus, it would not have been an interesting and useful feature.

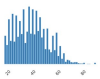


Figure 1: Distribution of Age

Figure 2: Distribution of Workclass

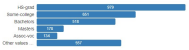


Figure 3: Distribution of Education

Figure 4: Distribution of Education number

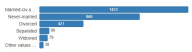


Figure 5: Distribution of Maritial status

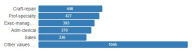


Figure 6: Distribution of Occupation

Figure 7: Distribution of Relationship

Figure 8: Distribution of Race

Figure 9: Distribution of Sex

Figure 10: Distribution of Capital Gain

Figure 11: Distribution of Capital Loss

Figure 12: Distribution of Hours per week

Figure 13: Distribution of Native Country

Figure 14: Distribution of Income (target)

Initially, the completeness of the dataset is 100%, and the dataset does not have any missing value. Five additional dataset has been created from the complete one: one having 10% of missing values, one having 20% of missing values, one having 30% of missing values, one having 40% of missing values, one having 50% of missing values. The missing values have been inserted completely at random.

3 Setup choices

3.1 Imputation techniques

3.1.1 Standard imputation

Exploring the variables, the values that each attribute assumes, and at the cardinality of each value, it is possible to choose how to treat each attribute for the imputation:

- *Age, education-num, hours-per-week* can be treated as a numeric value, and thus median imputation will be used (keeping it without decimal values)
- *Workclass, education, maritial-status, occupation, relationship, race, sex* are categorical. *Bfill, ffill* seems reasonable choices when we have few missing values, since it is a random imputation. The mode infact could change the distribution of the data a lot, when the data has lots of missing values. Thus, with 10% and 20% of missing values, the mode imputation seems a good choice, while with 30% and above of missing values a method such that *bbfill, ffill* that don't change the distribution are reasonable. For simplicity we will use *bfill* method
- *Capital-gain, Capital-loss* are numeric, but it is good to treat it as a categorical variable because they can assume few values. Since it seems that they have a particular value that is way more frequent than the others, a mode imputation seems reasonable for these variables.
- *Native country* is a categorical variable that has a particular value that is way more frequent than the others. A mode imputation seems reasonable for this variable.

3.1.2 Advanced imputation

The imputation will be done using MICE - Multiple Imputation by Chained Equations[1]. The idea is: we treat a feature that have missing values as a target variable where the missing data are the value to predict. Then, we use ML to predict these values. MICE starts from one variable that have missing values. For this first variable the missing values are imputed using some standard techniques. Once we have this feature without any missing value, we can consider a second variable that have missing values. This variable is considered as the target variable, and we will use ML to predict the missing values of this variable using as feature the variables that we imputed before. After having imputed the second variable, we can use this two variable to predict the missing values of a third variable. And so on for all the variable. At the end the missing values will have been imputed considering more information coming from the other variables.

Since there are few references on MICE with both categorical and numerical variables, a pure custom implementation has been made. Infact, at the beginning MICE has been applied with KNN on the numerical variables, imputing each variable based on the previous varibles already imputed. When it comes to the categorical variables, we consider each of these variables as target, we predict the missing values, we convert this variable in numeric using one-hot encoding, adn then we proceed with the next variables.

For what concern the types of the variables:

- *Age, education-num, hours-per-week, Capital-gain, Capital-loss* can be treated as a numeric value, and thus ML-based Imputation using KNN imputation will be used.
- *Workclass, education, maritial-status, occupation, relationship, race, sex, Native country* are categorical. We will use Decision Trees method.

3.2 Imputation accuracy assessment

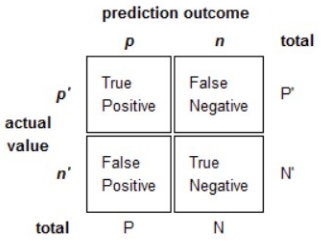
- The accuracy for the categorical variables will be the number of the correct imputed values over the number of elements of that variable.
- The accuracy for the numerical variables will be calculated as the cosine similarity between the imputed and the original one.

3.3 Machine learning algorithms

The target variable *income* contains two values. Thus the problem that can be faced by this dataset is classification. Logistic regression and Support vector machine will be used. To feed a ML algorithm with data, it should be with a specific structure. The first thing to notice is that the data consists of numerical and categorical data. Since Logistic regression and SVM will be used, they need the data to be numerical. Thus, hot-encode on the categorical variables has been applied, as well as a standardization using a z score: all the data will be standardized using the mean and the std dev.

3.4 Machine learning evaluation metrics

To evaluate the performances of the chosen method, we need to compute the confusion matrix which tells us the number of points which have been correctly classified and those which have been misclassified.



- Accuracy: $Acc = \frac{tp+tn}{N}$ fraction of the samples correctly classified in the dataset;
- Precision $Pre = \frac{tp}{tp+fp}$ fraction of samples correctly classified in the positive class among the ones classified in the positive class;
- Recall: $Rec = \frac{tp}{tp+fn}$ fraction of samples correctly classified in the positive class among the ones belonging to the positive class;
- F1 score: $F1 = \frac{2 \cdot Pre \cdot Rec}{Pre+Rec}$ harmonic mean of the precision and recall;

4 Results

4.1 Imputation accuracy

The first important result to discuss is the one related to the accuracy assessment of the imputed datasets.

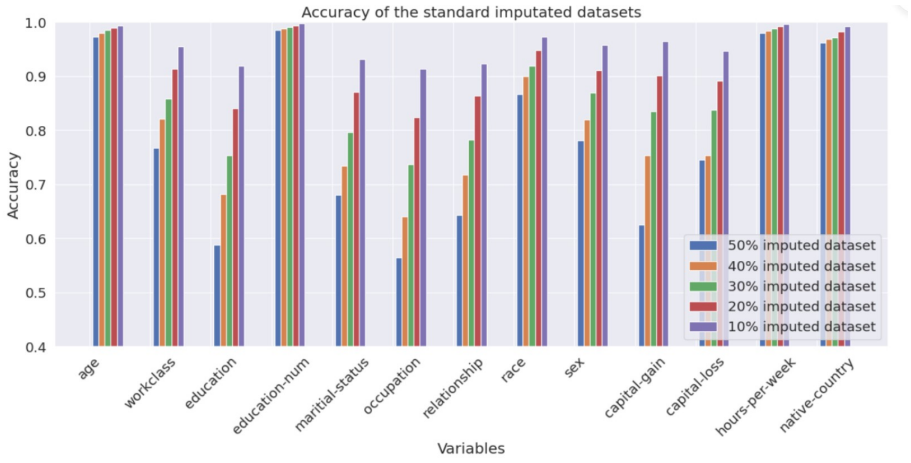


Figure 15: Accuracy of imputed data with standard imputation

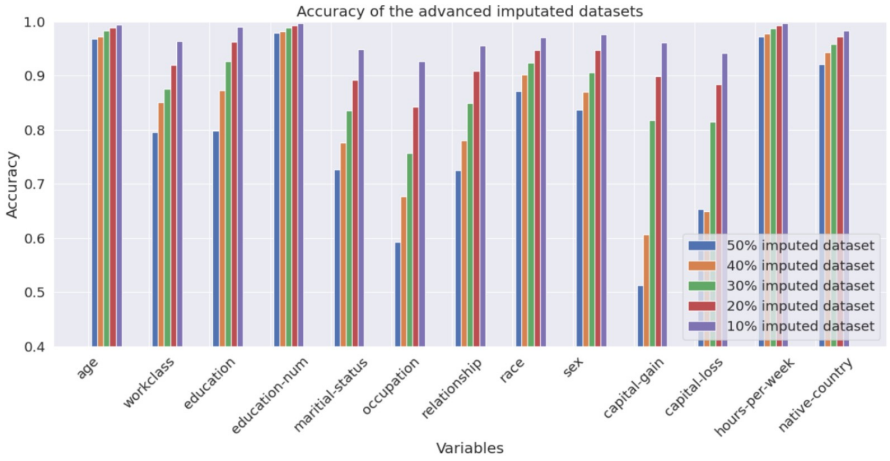


Figure 16: Accuracy of imputed data with advanced imputation

First, it is possible to note that the imputation on numerical variables is much more precise than the one in the categorical variables. This can be due to the nature of the variable: *age, education number, hours per week* are variables for which, even with a lot of missing value, it is possible to obtain imputed data that have more or less a similar distribution. For *capital gain, capital loss* that have a single value as the most frequent, imputation obtains the majority of value close to the most frequent one. It is different for the categorical variables. Infact, if the data has lots of missing values, it can happen that a specific value of that variable can lose all the observations, with a lack of information that can not be imputed anymore. This problem infact is solved in the numerical one because the predictions can be close to the real value, and also because the accuracy is measured using a similarity like the cosine. The only exception is *native country*, since it is a variable that has a value that is way the most frequent one, and also because during MICE it was the last categorical variable to be predicted, so it has all the other variables as features.

Another factor for the accuracy of the imputation can be the number of distinct values in the categorical variables.

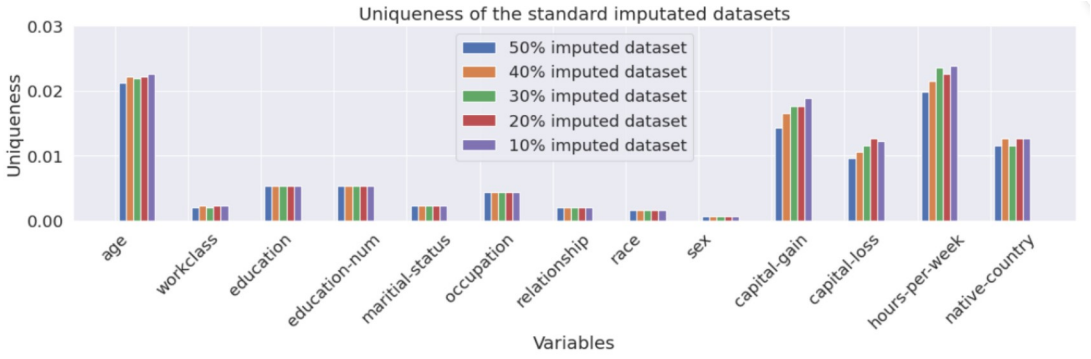


Figure 17: Uniqueness of imputed data with standard imputation

The less the distinct values, the closer are the imputed values to the original one. For example, occupation has more distinct values than the other categorical variables and the performance is worse than the others. Regarding the two importation methods, there is an important aspect. In general, the advanced imputation performs better than the standard, especially in education. We can see that in the capital gain and in the capital loss, the accuracy in the 50% missing dataset is lower than the standard one, because in the advanced the variable has been predicted numerically using the KNN, while in the standard we used the mode, and as a result, we kept the imputed values equal to the most frequent one. In general, the trend is that the less missing values are present in the dataset, the better is the similarity between imputed and original variables.

4.2 Machine learning algorithms performance

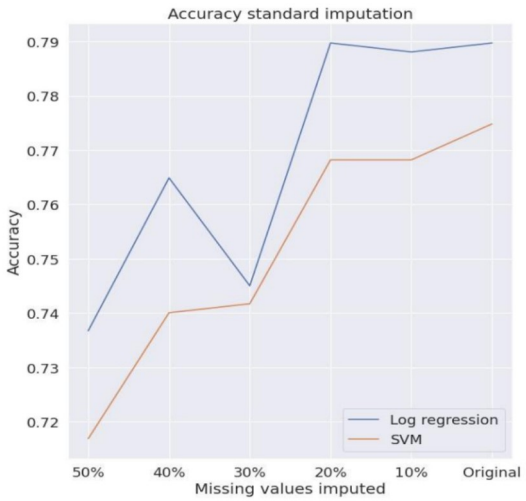


Figure 18: Accuracy of ML algorithms on imputed data with standard imputation

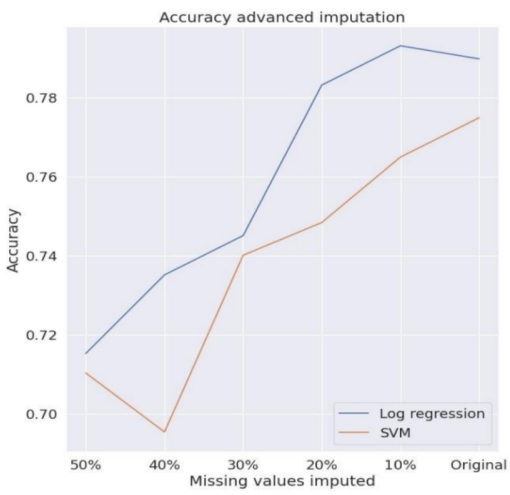


Figure 19: Accuracy of ML algorithms on imputed data with standard imputation

The first thing is that logistic regression is performing well and better than SVM. The reason can be multiple, and they depend on the tuning and on the optimization of them. Another aspect is that the logistic regression and the SVM have good performances in the standard imputation even with some missing value. The difference between standard and advanced is that the standard imputation is more unstable than the advanced one, since the latter has a more cleaned trend.

5 Conclusion

Imputation methods are different with each other, in terms of complexity and accuracy of the results. Standard techniques are easier to apply, but they impute missing value using a limited amount of information. Instead, advanced imputation are using ML and they are exploiting more information, imputing data using all the features present in the data. Also the performances of the ML algorithms

described some difference based on the data used, even if the project did not aim to improve the algorithms their selves.

References

[1] Sam Wilson. The mice algorithm. https://cran.r-project.org/package=miceRanger/vignettes/miceAlgorithm.html.