Journal of Statistical Modelling: Theory and Applications
Vol. 2, No. 2, 2021, pp. 103-112
Yazd University Press 2021

JSMTA

Research Paper

Missing data imputation using supervised learning methods

Behzad Rezaei Shiri, Samaneh Eftekhari Mahabadi*
School of Mathematics, Statistics and Computer Science, College of Science,
University of Tehran, Tehran 14155-6455, Iran

Received: September 30, 2020 / Revised: February 26, 2021 / Accepted: April 20, 2021

Abstract: Missing data is a very common problem in all research fields. Case deletion is a simple way to handle incomplete data sets which could mislead to biased statistical results. A more reliable approach to handle missing values is imputation which allows covariate-dependent missing mechanism, as well. This paper aims to prepare guidance for researchers facing missing data problems by comparing various imputation methods including machine learning techniques, to achieve better results in supervised learning tasks. A benchmark dataset has experimented and the results are compared by applying popular classifiers over varying missing mechanisms and rates on this benchmark dataset.

Keywords: Imputation; Machine learning algorithms; Missing data; Missing mechanism
Mathematics Subject Classification (2010): 68T09

1 Introduction

Missing data problem is a common issue in most real-world studies. Since most statistical models and data-dependent machine learning (ML) algorithms could only handle complete data sets, the issue of how to approach missing values plays an important role in statistical inferences.

Let Y be an $(N \times K)$ data matrix with i-th row $y_i = (y_{i1}, y_{i2}, ..., y_{iK})$ where y_{ij} is the value of j-th feature for the i-th sample. Define the subset of observed values as Y^{obs} and missing values as Y^{mis}. Also, let $M = [m_{ij}]$ be the missing indicator matrix, where m_{ij} indicates whether y_{ij} is missing or not.

*Corresponding author: seftekhari@ut.ac.ir

Rubin (19776) defines three different missing mechanisms according to the conditional probability of the missingness, $\{m_{ij} = 1\}$, given the data. The mechanism of missing data is completely at random (MCAR) if the probability of missingness is independent of all data values, missing or observed,

$$P(m_{ij} = 0|Y) = g(\phi), i = 1, ..., N, \quad j = 1, ..., K,$$

where $g(.)$ is a known link function and ϕ is the vector of unknown mechanism parameters. The missing mechanism is called missing at random (MAR) if the probability of missingness depends only on the observed data values,

$$P(m_{ij} = 0|Y) = g(Y^{obs}; \phi), i = 1, ..., N, \quad j = 1, ..., K.$$

Finally, the mechanism is called missing not at random (MNAR) when the probability of missingness may also depend on the unobserved data even after conditioning on the observed ones. The missing mechanism for the likelihood inferences is ignorable when the MCAR or MAR assumptions hold with the additional condition of disjoint parameter spaces of the missing mechanism and the data model (see Little and Rubin, 2014; Tsiatis, 2007, for more details).

One simple approach to analyze incomplete data is complete case (CC) analysis which discards all incomplete cases. This approach is logical only if the missing rate is considerably small or the missing data mechanism is MCAR (Little and Rubin, 2014). However, if the missing mechanism is MAR or MNAR or the missing rate is considerably high, the CC approach could highly influence statistical results. This is due to the fact that CC analysis makes no use of observed features of an incomplete case.

Suppose that y_{ij} is missing, but the value of another variable say y_{ik} is observed for the same sample which is highly correlated with y_{ij}, then it is tempting to predict the missing value of y_{ij} from y_{ik}, and then to include the filled-in (or imputed) value for further analyses. Such methods are called imputation methods which are general and flexible for handling missing-data problems. These methods can be applied to impute one value for each missing item (single imputation) or, to impute more than one value (multiple imputation), to allow appropriate assessment of imputation uncertainty.

Formally, imputations are means or draws of a suitable predictive distribution of the missing values given the observed ones. Thus, imputation procedure requires a method to create a predictive distribution based on the observed data. The predictive distribution might be defined based on a formal statistical model with explicit assumptions (such as linear or generalized linear models). On the other hand, the predictive distribution could focus on an algorithm, which implies an underlying model where the assumptions are implicit (such as neural networks algorithm), but they still need to be carefully assessed to ensure that they are reasonable.

This paper aims to examine three common ML classifiers: artificial neural networks (NNs), random forest (RF), and k-nearest neighbors (k-NN) for Incomplete train data due to missing categorical target variable. Actually, the performances of theses classification algorithms on imputed data set using various imputation methods under different missing scenarios are compared. The assessed imputation methods include mode imputation, random imputation and logistic regression as explicit ones and K-NN, classification and regression trees and random forests as implicit algorithms. The

results provide a guideline to help data analysts handling categorical missing data issues appropriately based on their research.

The paper is organized as follows; Section 2 explains the dataset and preprocessing applied for the analysis and missing data generating process. Section 3 involves ML algorithms employed for imputation. The numerical results of model training for the imputed datasets are given and their mean accuracies are compared for different ML algorithms and missing data scenarios in Section 4. Finally, Section 5 includes some concluding remarks.

2 Data description and preprocessing

The experimented data is excerpted from the 1994 Census benchmark database, named *Adult* (also, known as "Census Income" dataset)[†]. We have used a discretized features version of the dataset in which numerical variables are transformed into an ordinal scale. This dataset contains $N = 48,842$ individuals with $K = 15$ features. The 15-th feature indicates whether a person makes over 50000 dollars ($50K$) a year, which includes 11687 cases in the over $50K$ class, and the other features include some personal business information. The structure of this dataset and its features are shown in Table 1. Three features *workclass*, *occupation* and *native.country* contain 2799, 2809 and 2875 missing cases, respectively, with an overall rate of 7 percent and it seems that the missing mechanism is MNAR (see Kohavi, 1996; Jason and Paolo, 2018).

Table 1: structure of discretized adult dataset

age	workclass	fnlwgt	education	education.num
ordinal: 0-4	categorical: 8 categories	Integer>0	categorical: 16 categories	ordinal: 1-16
marital.status	occupation	relationship	race	sex
categorical: 7 categories	categorical: 14 categories	categorical: 16 categories	categorical: 5 categories	categorical: 2 categories
capitalgain	capitalloss	hoursperweek	native.country	class
ordinal: 0-4	ordinal: 0-4	ordinal: 0-4	categorical: 41 categories	categorical: 2 categories

In the pre processing phase, the *fnlwgt* column as the sampling weight (the number of people the census believes each record represents) which is not related to the target variable is omitted. Also, two variables *education* and *relationship*, which have the same information as *education.num* and *marital.status* respectively, are omitted. The *workclass* variable reclassified into three major classes, "self-employed", "government", and "other" to simplify and generalize our classifiers; also, *native.country* variable translated into two major classes, "USA" and "other". After removing missing observations, the dataset was randomly divided into the train and test sets, with proportions 2/3 and 1/3, respectively. The random partitioning is done using R random seed number 646 to be able to reproduce.

Then some missing values for the target variable (in the train sample) are generated according to four types of missing mechanism (MCAR, MAR, MNAR: stochastic right

[†]https://archive.ics.uci.edu/ml/datasets/adult

censoring, and MNAR: stochastic censoring). In the MCAR mechanism, probability of missingness in the target variable is assumed to be

$$p(m_i = 1|Y) = p(m_i = 1) = \delta; \forall i \in \{1, 2, ..., n\},$$

that is independent of either observed or unobserved data and n is the size of train sample. In the MAR mechanism it is assumed that

$$p(m_i = 1|Y) = p(m_i = 1|Y^{obs}) = exp\{\alpha + \beta age_i\}/(1 + exp\{\alpha + \beta age_i\}),$$

where the non-response probability of *income class* is assumed to be related to the completely observed feature, *age*. In the stochastic MNAR mechanism, it is assumed that

$$p(m_i = 1|Y) = p(m_i = 1|Y^{obs}, Y^{mis}) = exp\{\alpha_1 + \beta_1 class_i\}/(1 + exp\{\alpha_1 + \beta_1 class_i\}),$$

which means that the missingness stochastically could be dependent on the unobserved target value, *class*. Finally, in the stochastic right censoring mechanism, the following mechanism is assumed:

$$p(m_i = 1|Y) = \begin{cases} \delta & \text{if } class_i \in C \\ 0 & \text{if } o.w. \end{cases} \quad (1)$$

where C is "$>50K$" category.

After missing data simulation, numerical variables are normalized using min-max and Z-score normalizations, and all categorical variables except for the target variable (which contains only two classes) converted into indicator variables. It should be noticed that the target variable in the test sample is not changed during simulation procedure.

3 Imputation methods

Discarding observations with missing values and just using complete cases, wastes so much information and leads to poor results except when the missing rate is too low (e.g. lower than 5 percent) or the missing mechanism is MCAR (Graham, 2009; Schafer, 1999). However, imputation is a more reasonable way to treat missing data. Imputations based on explicit modeling assume that the missing mechanism is MAR Jason and Paolo (2018).

In related works, Batista and Monard (2003) applied k-NN imputation and decision trees (DTs) to substitute missing data. Also, Silva et al. (2011) compared the performance of imputation based on NNs with mode imputation for categorical variables, which shows that NNs achieve the best result. Van Buuren and Oudshoorn (1999) proposed a method to combine different regression models. Breiman (2001) presented Random Forests to deal with mixed-type missing data and Stekhoven and Buehlmann (2012) used an iterative process to impute multivariate missing data. Jason and Paolo (2018) compares classifiers trained on imputed data with different missing-data perturbation which shows that imputation methods can improve prediction accuracy by regularizing the classifier.

To impute perturbed train datasets, we tried seven imputation methods including three explicit methods, mode imputation, random imputation, and logistic regression (LR) and four ML imputation methods MissForest, K-NN , RF, and classification and regression trees (CART) algorithms. Also, the tuning parameters are chosen based on 5-fold cross-validation (CV).

We have also implemented a new ensemble learning imputation algorithm using a voting function that combines the results of the LR method and four ML imputation techniques and chooses the most voted level as our new imputation. Hence, we finally have multiple of eight imputation methods used to generate eight complete train data sets for different mechanisms and missing rates $(0.1, 0.25, 0.4)$.

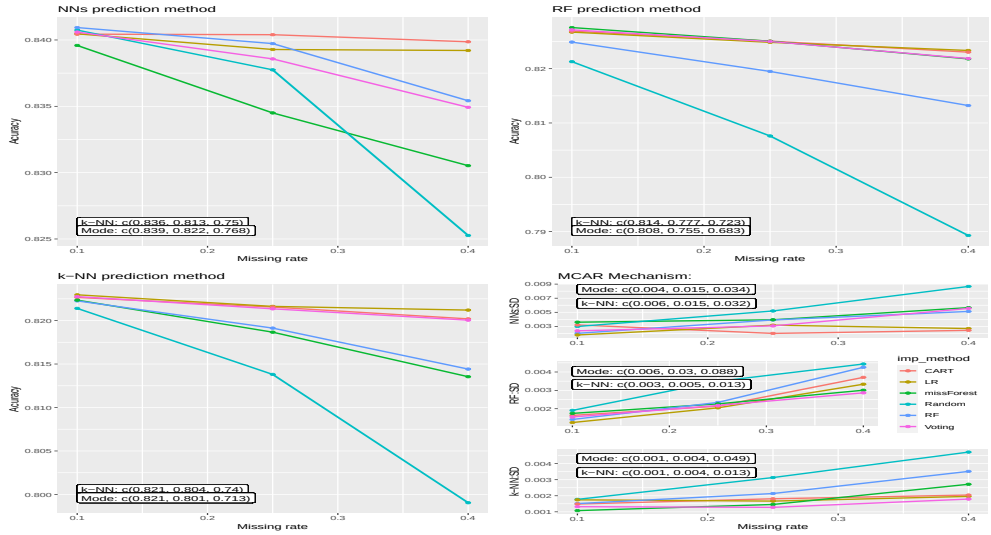

Figure 1: MCAR mechanism; using min-max normalization.

4 Results of numerical analysis

In this section, we will apply three well-known classifiers including Neural Networks, Random Forests and K-NN as the predictive machine learning algorithms. More specifically, NNs classifier is assumed to have two hidden layers, each containing 64 nodes with "relu" activation function, also the "sigmoid" activation function is used in the output layer. The optimization is done via the "adam" method using "binary cross-entropy" loss function trough 15 epochs. The RF algorithm was performed using $m = 150$ trees, and to perform k-NN classifier, $k = 32$ neighbors are assumed.

Firstly, the results of model fitting on the original dataset (before artificially missing values are inserted), shows that NNs, RF, and k-NN achieved 0.8400, 0.8347, and 0.8257 accuracies, respectively, using min-max normalization, and 0.8496, 0.8341, and 84.17 accuracy rates using Z-score normalization. It means that, NNs mean accuracies are the highest over both normalization methods among three assumed classifiers for the original data set.

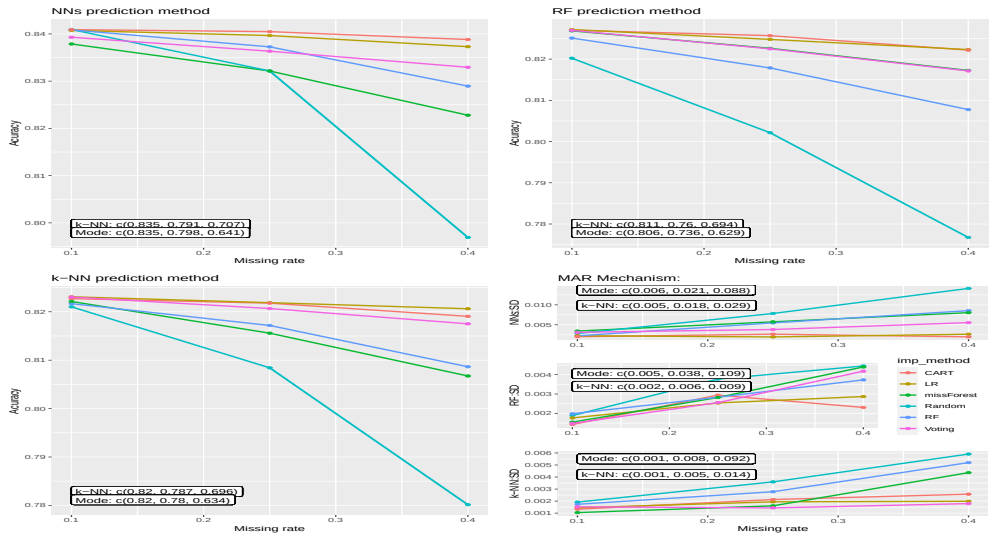

Figure 2: MAR mechanism; using min-max normalization.

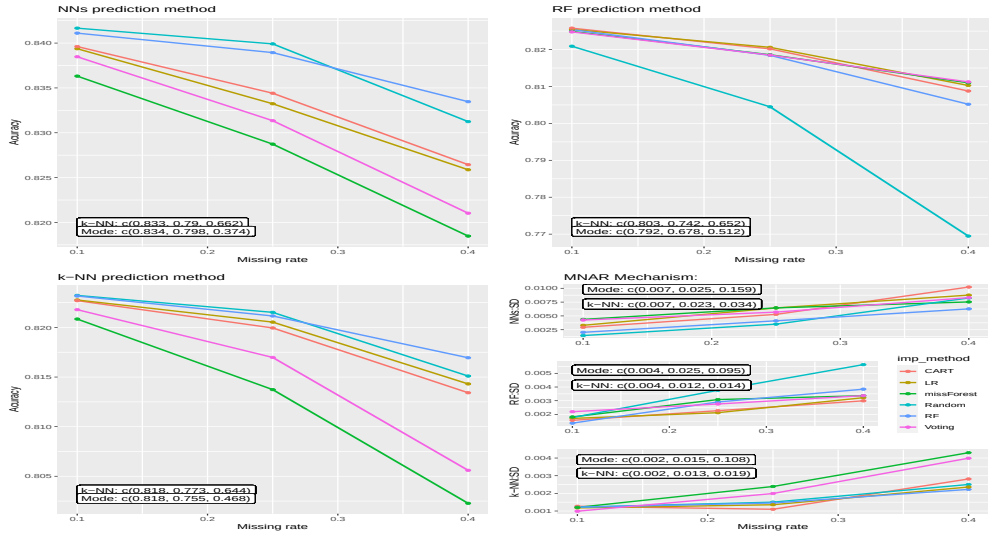

Figure 3: Stochastic MNAR mechanism; using min-max normalization.

To compare the performances of different imputation methods over different missing scenarios, these classifiers are fitted over imputed preprocessed training data sets and their accuracy rates on the test set (ignoring known target values) under different missing scenarios and two different transformations on the numerical features (min-max or Z-score) are calculated. To eliminate randomness effect, the fitting process is repeated 50 times in each scenario and the mean of accuracy rates with their standard deviations are calculated and plotted in Figures 1-4 and 5-8 for min-max and Z-score normalizations, respectively.

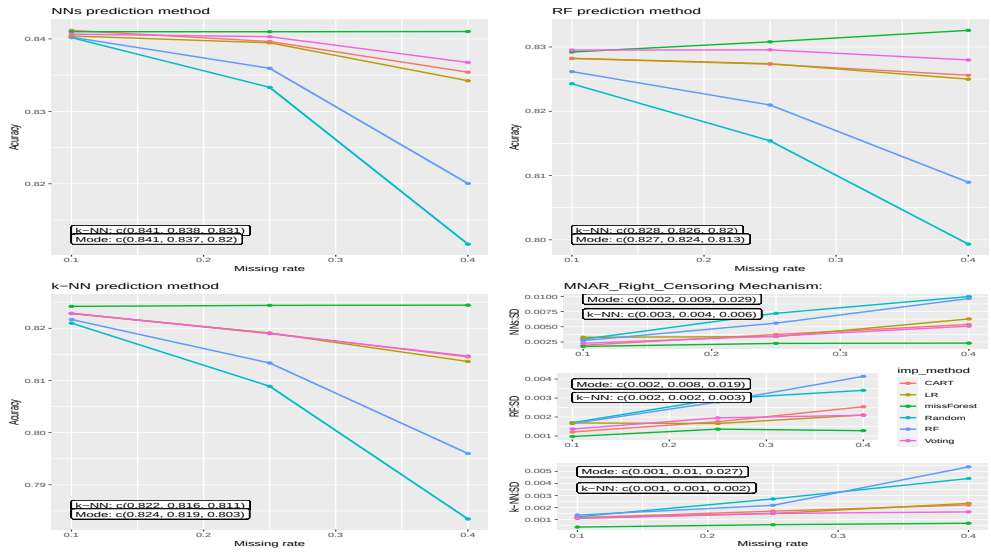

Figure 4: Stochastic MNAR mechanism; using min-max normalization.

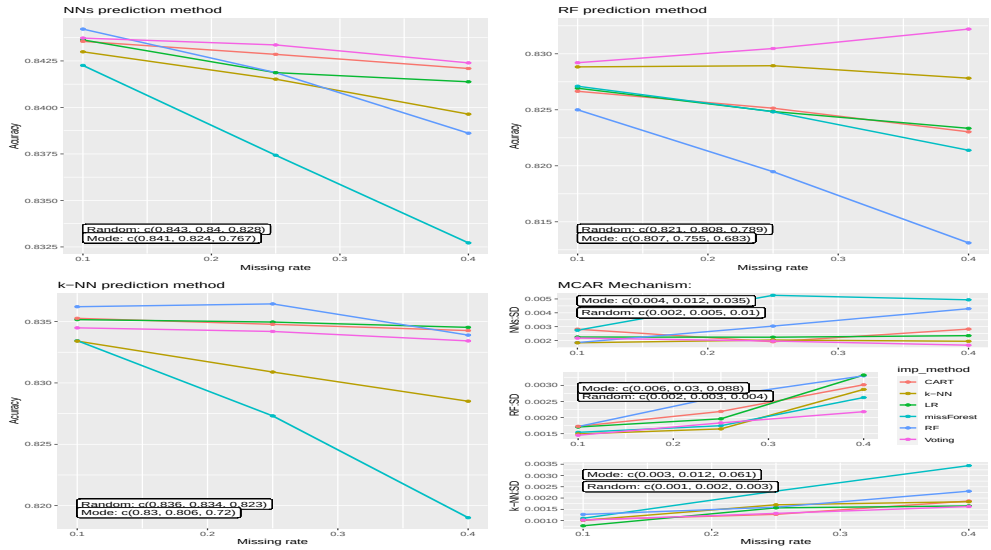

Figure 5: Average and standard deviation of accuracy rates of NNs, RF and K-NN classifiers under MCAR mechanism; using z-score normalization. In the upper panel and the bottem left panel, average accuracies are connected with lines for easier trace. Horizontal axis represent missing rate and the vertical axis represent mean accuracy of the corresponding classifier. Imputation methods are provided on the right side of every chart, note that since k-NN and Mode imputation results were far from the other values, they are reported numerically; c(.) stands for combination of results in three missing rates. Also, the standard deviation of accuracy rates are plotted in the bottem right panel.

First, we discuss imputation results based on the min-max transformation. Poorest accuracy results are produced by mode, and k-NN imputation methods. Unexpectedly,

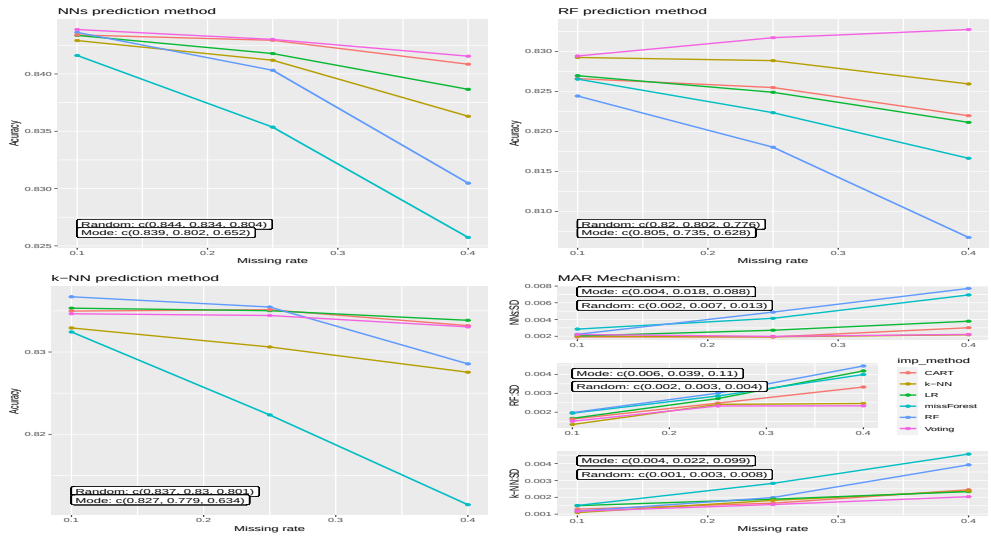

Figure 6: MAR mechanism; using z-score normalization.

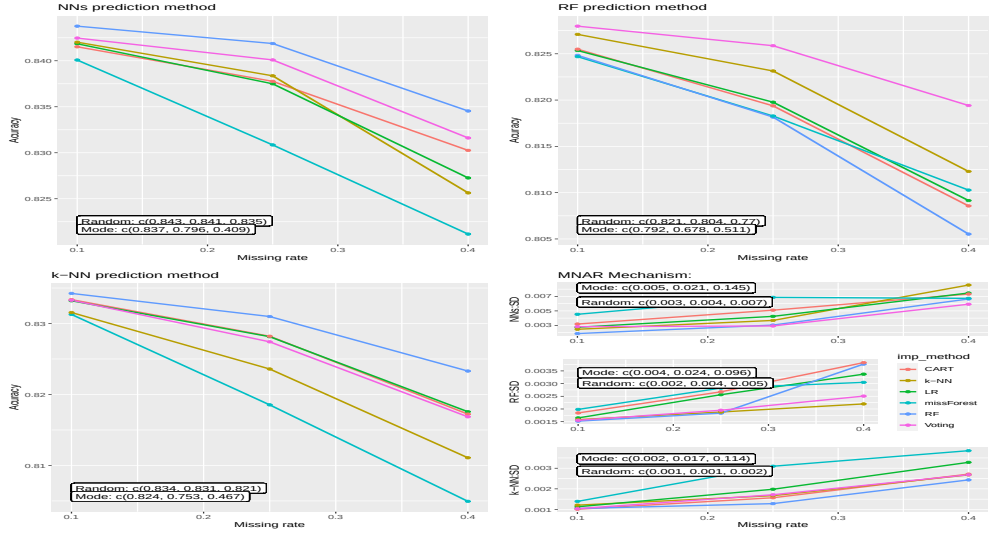

Figure 7: Stochastic MNAR mechanism; using z-score normalization.

random imputation performs better than k-NN imputation; one of the possible causes probably is the min-max normalizer used in the prepossessing step. In higher levels of missing rate, mode imputation leads to the poorest results and k-NN imputation also has very unsatisfying results. Under the MCAR mechanism with a low rate of missing rate ($\delta = 0.1$) results are very close and even perturbing the missing rates on this mechanism helps to improve prediction results. At higher rates of missing in the MCAR mechanism, CART imputation achieved the best overall results and prevented accuracy reduction in NNs classifier. Under MAR mechanism, results are similar to

MCAR and a lower missing rate leads to higher accuracies. Overall accuracies of CART and LR imputation methods are excellent even with a high rate ($\delta = 0.4$) of missing data. Under MNAR stochastic right censoring missForest leads to the best mean accuracies upon all prediction models. For the higher rates of missing, using our introduced voting imputation using RF predictive algorithm helped improving mean accuracy. Another noticeable result is in NNs classification is that random and RF imputation gained the best results on the MNAR mechanism and these prediction methods.

Finally, we assess results over Z-score imputed sets, in this case, mode and random imputation gained the lowest overall mean accuracy rates. Under the MCAR mechanism, voting imputation leads to best results using NNs and RF classifier, and even for higher rates of missing, accuracies of RF have improved. All other results are somewhat like under MCAR and voting imputation performs best overall. Under the MNAR mechanism, missForest is a good choice dealing with missing data using NNs and k-NN but voting imputation still gained the best mean accuracy using RF classifier. Finally, under MNAR stochastic right censoring, missForest imputation leads to the best overall mean accuracy.

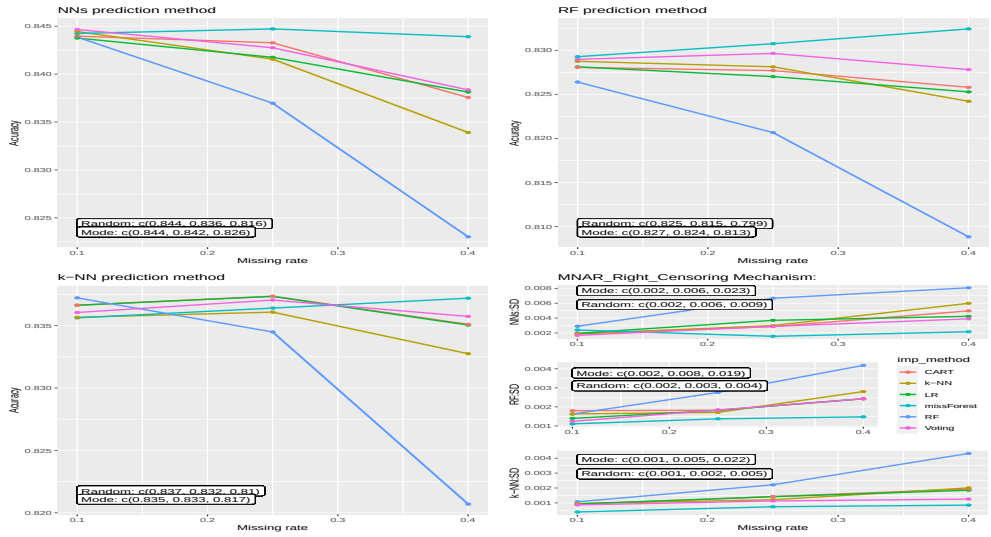

Figure 8: Stochastic MNAR mechanism; using z-score normalization.

Conclusions

In this paper, different imputation methods for incomplete binary target variable is explored. The evaluated imputation methods includes traditional mode and random methods, GLM-based imputation (binary logistic regression) and well known classification machine learning algorithms, random forest, K-NN and CART. Also, a benchmark dataset has been experimented to compare the performances of three different popular classifiers (NNs, RF, k-NN) on imputed train datasets, across different missing rates

and four mechanisms: MCAR, MAR, stochastic MNAR, MNAR right censoring. The numerical results show that CART imputation method achieved the best performance on these data. Actually, the proportion of missing data has the greatest effect on the results where the lower rates lead to improved prediction accuracies over different methods of imputation. Among implicit imputation methods, Mode imputation is generally the worst choice for imputation. As our results suggest when the missing mechanism is MCAR, ML imputation methods are appropriate choices. Specifically, CART imputation is a generally good method of imputing missing data which leads to the best performance under MCAR and MAR mechanisms. Our results suggest not to use Z-score normalization for numerical predictors and to avoid using mode imputation in a classification task with incomplete categorical target variable. Also, if the missing generating process is known to follow MNAR stochastic right censoring mechanism, it is best to use missForest imputation method.

References

Batista, G.E and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence,* **17**(5-6), 519–533.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Graham J.W. (2009) Missing data analysis: making it work in the real world. *Annual Review of Psychology,* **60**(1), 549–-576.

Poulos, J. and Valle, R. (2018). Missing data imputation for supervised learning. *Applied Artificial Intelligence*, **32**(2), 186–196.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifier: a Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* **96**(1), 202–207.

Little R. and Rubin D. (2014). *Statistical Analysis with Missing Data*. John Wiley and Sons.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, **63**, 581-592.

Schafer J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research* ,**8**(1), 3–-15.

Silva-Ramírez, E. L., Pino-Mejías, R., López-Coello, M., and Cubiles-de-la-Vega, M. D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks,* **24**(1), 121–129.

Stekhoven, D.J. and Buehlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data. *Bioinformatics,* **28**(1), 112–118.

Tsiatis A. (2007). *Semiparametric Theory and Missing Data*. Springer Science and Business Media.

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE* (pp. 1–20). Leiden: TNO.