No imputation without representation

Anonymous authors
Paper under double-blind review

Abstract

By filling in missing values in datasets, imputation allows these datasets to be used with algorithms that cannot handle missing values by themselves. However, missing values may in principle contribute useful information that is lost through imputation. The missing-indicator approach can be used in combination with imputation to instead represent this information as a part of the dataset. There are several theoretical considerations why missing-indicators may or may not be beneficial, but there has not been any large-scale practical experiment on real-life datasets to test this question for machine learning predictions. We perform this experiment for three imputation strategies and a range of different classification algorithms, on the basis of twenty real-life datasets. We find that on these datasets, missing-indicators generally increase classification performance. In addition, we find no evidence for most algorithms that nearest neighbour and iterative imputation lead to better performance than simple mean/mode imputation. Therefore, we recommend the use of missing-indicators with mean/mode imputation as a safe default, with the caveat that for decision trees, pruning is necessary to prevent overfitting. In a follow-up experiment, we determine attribute-specific missingness thresholds for each classifier above which missing-indicators are more likely than not to increase classification performance, and observe that these thresholds are much lower for categorical than for numerical attributes. Finally, we argue that mean imputation of numerical attributes may preserve some of the information from missing values, and we show that in the absence of missing-indicators, it can similarly be useful to apply mean imputation to one-hot encoded categorical attributes instead of mode imputation.

1 Introduction

Missing values are a frequent issue in real-life datasets, and the subject of a large body of ongoing research. Apart from a diverse range of proposals specific to individual algorithms, a common general strategy is to replace the missing value with an estimated value: imputation. An advantage of imputation is that we obtain a complete dataset, to which we can apply any and all algorithms that make no special provision for missing values. However, missing values may be informative, and a disadvantage of imputation is that it eliminates this information.

The missing-indicator approach (Cohen, 1968) is a well-established way to represent and thereby preserve the information encoded by missing values. For every original attribute, this approach adds a new binary 'indicator' or 'dummy' attribute that takes a value of 1 if the value for the original attribute is missing, and 0 if not.[1] The missing-indicator approach is often presented as an alternative to imputation, but since it does not resolve the missing values in the original attributes, it can only be used in addition to, not instead of imputation.

Both imputation and the missing-indicator approach originate in the statistical literature. While imputation strategies have been the subject of a rich body of research, the missing-indicator approach has not received a large amount of attention, and is often dismissed or disregarded in overviews of approaches towards missing values. It is therefore an open question whether missing-indicators should be used for predictive tasks in

[1] Some authors use the opposite convention, letting the indicator express non-missingness.

machine learning (Sperrin et al., 2020). On the one hand, the addition of missing-indicators results in a more complete, higher-dimensional representation of the data. On the other hand, their omission can be seen as a form of dimensionality reduction, which may increase the efficiency and effectiveness of a dataset by eliminating redundancy.

To determine whether this trade-off is useful, a key question is to which extent missing values in a given dataset are informative, and to which extent their distribution is predictable. In the latter case, the phrase "missing at random" (MAR) (Rubin, 1976) is used to indicate that the distribution of missing values is dependent on the known values, while the stricter phrase "missing completely at random" (MCAR) denotes values that are distributed truly randomly. For real-life datasets, unless we have specific knowledge about the process responsible for the missing values, we have to assume some degree of informativeness in principle.[2] However, Schafer (1997) has argued that in practice, the attributes of a dataset can be sufficiently redundant that one can get away with assuming its missing values are MAR. But even if this is so, imputation may not always perform optimally, in which case missing-indicators may still prove useful.

In light of these conflicting arguments, the usefulness of missing-indicators in machine learning is ultimately an empirical question. However, previous experiments in this direction have been limited in scope and number. These limitations include the use of only one or a handful of datasets, the use of datasets from which values have been removed artificially, at random (corresponding to the MCAR setting), and not comparing the same imputation strategies with and without missing-indicators.

The purpose of the present paper is straightforward. On the basis of twenty real-life classification problems with naturally occurring missing values, we evaluate the usefulness of missing-indicators for three common types of imputation and a range of popular classification algorithms.

In Section 2, we provide a brief overview of the existing literature on missing-indicators, including previous experimental evaluations. In Section 3, we describe our experimental setup. We report our results in Section 4 and conclude in Section 5.

2 Background

We start with a brief discussion of the origins and reception of the missing-indicator approach, as well as previous experimental evaluations of the use of missing-indicators in prediction tasks.

2.1 Origins and reception

The missing-indicator approach originates in the statistical literature on linear regression, and dates back to at least Cohen (1968). Cohen pointed out that values in real-life datasets are typically not missing completely at random, and that the distribution of missing values may in particular depend on the values of the attribute that is to be predicted. He proposed that each attribute could be said to have two 'aspects', its value, and whether that value is present to begin with, which should be encoded with a pair of variables. For missing attribute values, the first of these variables was to be filled in with the mean of the known values, although other applications might call for different values. Cohen's proposal was subsequently expanded in Cohen & Cohen (1975), but received only limited recognition in the following years (Kim & Curry, 1977; Stumpf, 1978; Chow, 1979; Hutcheson & Prather, 1981; Anderson et al., 1983; Orme & Reis, 1991).

Cohen's proposal was subjected to a formal analysis by Jones (1996), who showed that, if one assumes that missing values are MAR, and the true linear regression model does not contain any terms related to missingness, it produces biased estimates of the regression coefficients (unless the sample covariance between independent variables is zero). However, these assumptions run directly counter to the position set out by Cohen & Cohen (1975) that a priori, the missingness of each attribute is a possible explanatory factor, that it is safer not to assume that missing values are distributed randomly, and that the usefulness of missing-indicators is ultimately an empirical question.

[2] This is acknowledged by authors working under the assumption of MAR, e.g. "When data are missing for reasons beyond the investigator's control, one can never be certain whether MAR holds. The MAR hypothesis in such datasets cannot be formally tested unless the missing values, or at least a sample of them, are available from an external source." (Schafer, 1997)

Allison (2001), motivated by Jones (1996) and working under the general assumption of MAR, dismissed missing-indicators as "clearly unacceptable", before conceding that they in fact produce optimal estimates when the missing value is not just missing, but cannot exist, such as the marital quality of an unmarried couple. However, this semantic distinction may not always be clear-cut in practice, and the more pertinent question may be whether missing values are informative. Allison (2010) later acknowledged that missing-indicators may lead to better predictions and their use for that purpose was acceptable. Missing-indicators have also been dismissed by Pigott (2001); Schafer & Graham (2002); Graham (2009); Aste et al. (2015), and are frequently omitted in overviews of missing data strategies (Schafer, 1997; Enders, 2010; Eirola, 2014; García et al., 2015; Das et al., 2018).

2.2 Previous experiments

Only a handful of experimental comparisons of missing data approaches have included the missing-indicator approach, and these have been limited in scope. Vamplew & Adams (1992) and Ng & Yusoff (2011) only use a single dataset with randomly removed values, and base their evaluation on the performance of a single algorithm (respectively a neural network and linear regression). Pereira Barata et al. (2019) use three classification algorithms and 22 datasets, but again with randomly removed values, explicitly assuming a MCAR context. They conclude that imputation outperforms missing-indicators, but the comparison is not like-for-like, since it involves several forms of imputation but only combines indicator attributes with zero imputation. Van der Heijden et al. (2006) compare missing-indicators with zero imputation against several other forms of imputation without missing-indicators on one real dataset, for logistic regression. However, it appears that they do not use a test set, and only evaluate the resulting models on the training set.

Ding & Simonoff (2010) conduct a more extensive investigation, using insights from a series of Monte Carlo simulations to systematically remove values from 36 datasets to simulate different forms of missingness. They use these datasets to compare zero imputation[3] with indicator attributes against mean/mode imputation without, as well as a number of other missing data approaches, for logistic regression. In addition, the authors evaluate a related representation of missing values[4] on the same set of 36 datasets, and on one real-life dataset with missing values, for decision trees. They find that there is strong evidence that representing missing values is the best approach when they are informative; when this is not the case their results show no strong difference with respect to imputation.

The comparison by Grzymala-Busse & Hu (2000) is based on 10 datasets with naturally occurring missing values. However, the setting is purely categorical — all attributes are transformed into categorical attributes, the only form of imputation is mode imputation, and the missing value approaches are evaluated on the basis of the LERS classifier (Learning from Examples based on Rough Sets (Grzymala-Busse, 1988)).

Marlin (2008) compares zero imputation with missing-indicators (*augmentation with response indicators*) against several forms of imputation without, for logistic regression and neural networks, on the basis of an extensive series of simulations, one dataset with artificially removed values, and three real datasets. For the real datasets, there is no strong difference in performance between the different approaches.

Finally, and most recently, Perez-Lebel et al. (2022) compare four different imputation techniques with and without missing-indicators (*missingness mask*) on seven prediction tasks derived from four real medical datasets, and conclude that missing-indicators consistently improve performance for gradient boosted trees, ridge regression and logistic regression.

Even experimental comparisons of missing data that do not feature the missing-indicator approach generally do not involve more than a handful of real-life datasets with naturally occurring missing values. We have only found Luengo et al. (2012a;b), who use 21 datasets from the UCI repository, but 12 of these are problematic.[5]

[3]Presumably, Ding & Simonoff (2010) use one-hot encoding for categorical attributes, in which case zero imputation is equivalent to treating missing values as a separate category, but they do not state this explicitly.

[4]For categorical values, encoding missing values as a separate category, for numerical values, encoding missing values as an extremely large value that can always be split from the other values.

[5]The target column of the *echocardiogram* dataset ('alive-at-1') is supposed to denote whether a patient survived for at least one year, but it doesn't appear to agree with the columns from which it is derived, that denote how long a patient (has) survived and whether they were alive at the end of that period. The *audiology* dataset has a large number of small classes with

Table 1: Classification Algorithms

Name	Description
NN-1	Nearest neighbours (Fix & Hodges, 1951) with (Boscovich) 1-distance
NN-2	Nearest neighbours with (Euclidean) 2-distance
NN-1-D	Nearest neighbours with 1-distance, distance-weighted (Dudani, 1976)
NN-2-D	Nearest neighbours with 2-distance, distance-weighted
SVM-L	Soft-margin Support Vector Machine (Cortes & Vapnik, 1995) with linear kernel
SVM-G	Soft-margin Support Vector Machine with Gaussian kernel
LR	Multinomial logistic regression (Cox, 1966)
MLP	Multilayer perceptron (Rosenblatt, 1961) with ReLu activation (Fukushima, 1969), Glorot initialisation (Glorot & Bengio, 2010) and Adam optimisation (Kingma & Ba, 2014)
CART	Classification and Regression Tree (Breiman et al., 1984)
RF	Random Forest (Breiman, 2001)
ERT	Extremely Randomised Trees (Geurts et al., 2006)
ABT	Ada-boosted trees (Freund & Schapire, 1995) with SAMME (stagewise additive modeling using a multi-class exponential loss function) (Zhu et al., 2009)
GBM	Gradient Boosting Machine (Friedman, 2001)

3 Experimental setup

To evaluate the effect of the missing-indicator approach on classification performance, we conduct a series of experiments, using the Python machine learning library *scikit-learn* (Pedregosa et al., 2011).

3.1 Evaluation

We preprocess datasets by standardising numerical attributes and one-hot encoding categorical attributes (as required by the implementations in scikit-learn).

We measure classification performance by performing stratified five-fold cross-validation, repeating this for five different random states (which determine both the dataset splits and the initialisation of algorithms with a random component), and calculating the mean area under the receiver operator curve (AUROC). For multi-class datasets, we use the extension of AUROC defined by Hand & Till (2001).

To compare two alternatives A and B, we consider the p-value of a one-sided Wilcoxon signed-rank test (Wilcoxon, 1945) on the mean AUROC scores for our selection of datasets. When we compare A vs B, a score below 0.5 means that A increased performance on our selection of datasets; the lower the scores, the more confident we can be that this generalises to other similar datasets. Conversely, a score higher than 0.5 means that A decreased performance on our selection of datasets.

3.2 Imputation Strategies

We consider the following three imputation strategies:

- *Mean/mode imputation* replaces missing values of numerical and categorical attributes by, respectively, the mean and the mode of the non-missing values.

complex labels and should perhaps be analysed with multi-label classification. In addition, it has ordinal attributes where the order of the values is not entirely clear, and three different values that potentially denote missingness ('?', 'unmeasured' and 'absent'), and it is not completely clear how they relate to each other. The *house-votes-84* dataset contains '?' values, but its documentation explicitly states that these values are not unknown, but indicate different forms of abstention. The *ozone* dataset is a time-series problem, while the task associated with the *sponge* and *water-treatment* datasets is clustering, with no obvious target for classification among their respective attributes. Finally, the *breast-cancer* (9), *cleveland* (7), *dermatology* (8), *lung-cancer* (5), *post-operative* (3) and *wisconsin* (16) datasets contain only very few missing values, and any performance difference between missing value approaches on these datasets may to a large extent be coincidental.

Table 2: Real-life classification datasets with missing values from the UCI repository for machine learning

Dataset	Records	Classes	Attributes			Missing value rate			Source
			Num	Cat	Total	Num	Cat	Total	
adult	48 842	2	5	8	13	0.0	0.017	0.010	Kohavi (1996)
agaricus-lepiota	8124	2	2	20	22	0.0	0.015	0.014	Schlimmer (1987)
aps-failure	76 000	2	170	0	170	0.083		0.083	Ferreira Costa & Nascimento (2016)
arrhythmia	452	13	279	0	279	0.0032		0.0032	Güvenir et al. (1997)
bands	540	2	19	15	34	0.054	0.054	0.054	Evans & Fisher (1994)
ckd	400	2	24	0	24	0.11		0.11	Rubini & Eswaran (2015)
crx	690	2	6	9	15	0.0060	0.0068	0.0065	Quinlan (1987)
dress-sales	500	2	3	9	12	0.20	0.19	0.19	
exasens	399	4	7	0	7	0.43		0.43	Soltani Zarrin et al. (2020)
hcc	165	2	49	0	49	0.10		0.10	Santos et al. (2015)
heart-disease	1611	2	13	1	14	0.18	0.0	0.17	Detrano et al. (1989)
hepatitis	155	2	19	0	19	0.057		0.057	Efron & Gong (1981)
horse-colic	368	2	19	1	20	0.25	0.39	0.26	McLeish & Cecile (1990)
mammographic-masses	961	2	2	2	4	0.042	0.041	0.042	Elter et al. (2007)
mi	1700	8	111	0	111	0.085		0.085	Golovenkin et al. (2020)
nomao	34 465	2	89	29	118	0.38	0.37	0.38	Candillier & Lemaire (2012)
primary-tumor	339	21	16	1	17	0.029	0.20	0.039	Cestnik et al. (1987)
secom	1567	2	590	0	590	0.045		0.045	McCann et al. (2008)
soybean	683	19	22	13	35	0.099	0.096	0.098	Michalski & Chilausky (1980)
thyroid0387	9172	18	7	22	29	0.22	0.0015	0.055	Quinlan et al. (1986)

- *Nearest neighbour imputation* (Troyanskaya et al., 2001) replaces missing values of numerical and categorical attributes by, respectively, the mean and the mode of the 5 nearest non-missing values, with distance determined by the corresponding non-missing values for the other attributes.
- *Iterative imputation*, as implemented in scikit-learn, based on Van Buuren & Groothuis-Oudshoorn (2011), predicts missing values of one attribute on the basis of the other attribute values using a round-robin approach. For numerical attributes, this uses Bayesian ridge regression (Tipping, 2001), initialised with mean imputation, while for categorical attributes, we use logistic regression, initialised with mode imputation.

The scikit-learn implementations of nearest neighbour and iterative imputation can currently only impute numerical features, so we had to adapt them for categorical imputation. In all other aspects, we follow the default settings of scikit-learn.[6]

3.3 Classification Algorithms

We consider the classification algorithms listed in Table 1, as implemented in scikit-learn. Hyperparameters take their default values, except for SVM-L, LR and MLP, where we increase the maximum number of iterations to 10 000 to increase the probability of convergence.

3.4 Datasets

We use twenty real-life datasets with naturally occurring missing values from the UCI repository for machine learning (Dua & Graff, 2019) (Table 2). We have preprocessed these datasets in the following manner. When it was clear from the description that an attribute was categorical, we treated it as such, even if it was originally represented with numerals. Conversely, where the possible values of an attribute admitted a semantic order, we encoded them numerically. We have removed attributes that were labelled non-informative by the accompanying documentation, as well as identifiers and alternative target values. For *dress-sales* (which does not appear to have been used in any publication), we cleaned up the data by eliminating spelling variations.

[6]For the *nomao* dataset, iterative imputation diverged, so we had to restrict imputation to the interval $[-100, 100]$.

Table 3: One-sided p-values, imputation with missing-indicators vs without

Classifier	Imputation strategy		
	Mean/mode	Neighbours	Iterative
NN-1	0.024	0.0027	0.0011
NN-2	0.035	0.0050	0.00085
NN-1-D	0.016	0.0031	0.00085
NN-2-D	0.0063	0.0070	0.00042
SVM-L	0.18	0.31	0.11
SVM-G	0.0063	0.0063	0.0027
LR	0.092	0.079	0.074
MLP	0.0050	0.013	0.011
CART	0.84	0.75	0.70
RF	0.058	0.12	0.29
ERT	0.36	0.018	0.027
ABT	0.089	0.10	0.49
GBM	0.39	0.022	0.18

Table 4: One-sided p-values, missing-indicators with iterative and nearest neighbour vs mean/mode imputation

Classifier	Imputation strategy	
	Neighbours	Iterative
NN-1	0.90	0.27
NN-2	0.74	0.26
NN-1-D	0.95	0.71
NN-2-D	0.80	0.34
SVM-L	0.48	0.61
SVM-G	0.47	0.94
LR	0.36	0.85
MLP	0.29	0.56
CART	0.67	0.69
RF	0.63	0.86
ERT	0.47	0.51
ABT	0.63	0.94
GBM	0.94	0.83

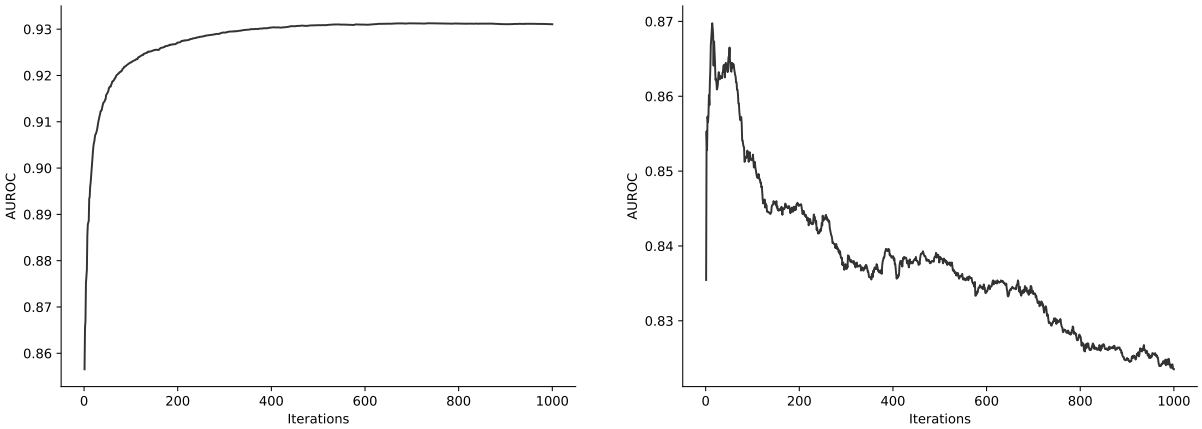
For *heart-disease*, we reduced the id-attribute to only identify the source hospital. In *thyroid0387*, a small number of records belonged to multiple classes, which we resolved by retaining the most specific class.

4 Results

Using the experimental setup detailed in the previous section, we now present the results of our evaluation of the missing-indicator approach.

4.1 Main results

The p-values obtained by comparing imputation with and without missing-indicators are displayed in Table 3. Missing-indicators generally lead to increased performance — with the notable exception of CART, to which we return below.



(a) adult

(b) mammographic-masses

Figure 1: Test AUROC for GBM for two illustrative datasets, using mean/mode imputation without missing-indicators, for one random state and one cross-validation fold.

Next, we consider how useful the more complicated imputation strategies are with respect to mean/mode imputation when we pair imputation with missing-indicators. Not very much, it turns out (Table 4). At best, nearest neighbour and iterative imputation only lead to a modest improvement, and for many classifiers, they actually decrease performance. Therefore, we focus on mean/mode imputation for the remainder of this section.

A possible reason for the poor performance of missing-indicators with CART, is that by default, the scikit-learn implementation of this classifier does not perform pruning, making it prone to overfitting. To test this hypothesis, we repeat our experiment for CART and mean imputation, but this time we apply cost complexity pruning ($\alpha = 0.01$). The resulting AUROC scores are now much better than the original scores without missing-indicators ($p = 0.013$) and somewhat better than cost complexity pruning without missing-indicators ($p = 0.23$).

In the case of ERT, missing-indicators may not lead to a clear performance improvement because of under-fitting. If we increase the number of trees from the default 100 to 1000, the improvement becomes somewhat clearer ($p = 0.15$).

For GBM, the default choice of 100 iterations of gradient descent can lead to both under- or overfitting, depending on the dataset (Fig. 1). We believe that it is generally preferable to continue training until an early-stopping criterion is met. If we apply the same criterion as with MLP,[7] the performance increase due to missing-indicators also becomes clearer ($p = 0.19$).

4.2 Missing thresholds

The theoretical motivation for representing missing values through missing-indicators is that this allows classifiers to learn the information encoded in their distribution. In general, this should be easier when there are more examples to learn from. If true, we can use this to obtain a better understanding of when missing-indicators might be useful on a per-attribute level.

We test this with the following additional experiment. For each attribute with missing values in each dataset, we reduce the original dataset by removing all other attributes with missing values. We thus obtain 1148 derived datasets, on which we again apply each of our classifiers (with pruning for CART, 1000 trees for ERT and early-stopping for GBM) and consider whether missing-indicators increase or decrease AUROC

[7]Setting aside 10% of the data for validation, stopping when validation loss has not decreased by at least 0.0001 for ten iterations, with a maximum of 10 000 iterations.

Table 5: Thresholds above which missing-indicators are more likely than not to increase AUROC, in terms of the absolute number of missing values or the missing rate

Classifier	Missing values		Missing rate	
	Categorical	Numerical	Categorical	Numerical
NN-1	20	296		
NN-2	10	146		
NN-1-D	21	370		
NN-2-D	6	80		
SVM-L			0.0	0.0
SVM-G			0.0	0.44
LR			0.0	0.0
CART	1	13		
ERT			0.0	1.0
ABT	1	18850		
GBM			0.0	0.0

(we dismiss ties). Finally, for each classifier we fit a logistic regression model with cluster robust covariance (clustered by the originating dataset), with the following potential parameters: categoricalness (whether the attribute is categorical) and either the number of missing values (log-transformed) or the missing rate. We use the Akaike information criterion (Akaike, 1971) to decide whether to select these parameters.

We find that for most classifiers, either the absolute or the relative number of missing values is an informative parameter with positive coefficient. For MLP, neither parameter is informative, while for RF, the number of missing values is an informative parameter with negative coefficient, for which we have no explanation at present. For every classifier except NN-1, NN-1-D and LR, categoricalness is an informative parameter with positive coefficient, meaning that missing-indicators are more beneficial for categorical than for numerical attributes.

The fitted logistic regression models allow us to calculate attribute-specific thresholds above which missing-indicators are more likely than not to increase AUROC, for all classifiers except MLP and RF (Table 5). In many cases, these thresholds are 1 or 0.0, indicating that missing-indicators are always likely to increase AUROC. We have included NN-1 and NN-1-D in this table on the basis of a model that includes the categoricalness parameter, since we find it implausible that it should not be relevant only for these specific classifiers. If we exclude it, the thresholds respectively become 233 and 285 records for categorical and numerical attributes alike. For LR, the threshold is a missing rate of 0.0, whether we include the categoricalness parameter or not.

4.3 Numerical imputation of missing categorical values

As indicated above, missing-indicators are generally more likely to increase performance for categorical than for numerical attributes. A potential explanation for this is the fact that the mode of a categorical attribute is one of the non-missing values, whereas the mean of a numerical attribute is generally not equal to one of the non-missing values. Therefore, categorical imputation renders missing values truly indistinguishable from non-missing values, whereas numerical imputation does not — the information expressed by missing values may be partially recoverable. This is particularly evident in the case of numerically encoded binary attributes, where mean imputation will represent missing values as a third, intermediary value.

We can achieve a similar partial representation of missing categorical values by changing the order in which we perform imputation and one-hot encoding, i.e. by performing numerical imputation on one-hot encoded categorical attributes with missing values. For imputation without missing-indicators, this indeed leads to

Table 6: One-sided p-values, mean imputation after one-hot encoding vs mode imputation of missing categorical values.

Classifier	Without missing-indicators	With missing-indicators
NN-1	0.030	0.22
NN-2	0.32	0.17
NN-1-D	0.030	0.36
NN-2-D	0.29	0.17
SVM-L	0.44	0.71
SVM-G	0.22	0.56
LR	0.88	0.023
MLP	0.14	0.52
CART	0.50	0.34
RF	0.084	0.78
ERT	0.023	0.95
ABT	0.56	0.66
GBM	0.12	0.56

better performance for some classifiers, while in combination with missing-indicators, it does not make much of a difference (Table 6)[8].

We finish this section with a note of caution: because we have performed multiple statistical tests, it is quite likely that some particular results do not generalise to other datasets, without necessarily changing the overall picture. It is difficult to say much more about this, because the tests are not independent of each other.

5 Conclusion

We have presented the first large-scale experimental evaluation of the effect of the missing-indicator approach on classification performance, conducted on real datasets with naturally occurring missing values, paired with three different imputation techniques. The central question was whether, on balance, more benefit can be derived from the additional information encoded in a representation of missing values, or from the lower-dimensional projection of the data obtained by omitting missing-indicators.

We found that, on the whole, missing-indicators increase performance for the classification algorithms that we considered, although we cannot be sure for each classifier that this result will generalise to other datasets. The only classifier for which missing-indicators decreased performance was CART. We showed that this is due to overfitting by the default configuration of the scikit-learn implementation of CART, and that missing-indicators do increase performance when pruning is applied. For ERT and GBM, we were able to show that the advantage of including missing-indicators becomes more significant when the number of trees of ERT is increased to limit underfitting, and the number of iterations of GBM is determined dynamically by an early-stopping criterion to avoid both under- and overfitting.

We also looked at the relative performance of the three different imputation strategies and found that, in the presence of missing-indicators, nearest neighbour and iterative imputation do not increase performance over simple mean/mode imputation, with the possible exception of NN-2 and NN-2-D in the case of iterative imputation. This is a useful finding, because implementations of more sophisticated imputation strategies may not always be available to practitioners working in different frameworks, or easy to apply.

In a follow-up experiment, we were able to determine attribute-specific missingness thresholds above which missing-indicators are more likely than not to increase performance. We found that for categorical attributes,

[8]LR is an exception here. We have no explanation for this, although we note that it corresponds with our finding in Subsection 4.2 that categoricalness is not a relevant factor for LR.

this threshold is generally very low, while for numerical attributes, there is more variation among classifiers, in particular as to whether this threshold is absolute or relative to the total number of records.

A possible explanation for the finding that missing-indicators are more useful for categorical than for numerical attributes is the fact that the mean of a numerical attribute is not generally identical to any of the non-missing values, and that mean imputation therefore preserves some of the information of missing values. This is supported by the results of a further experiment, which showed that in the absence of missing-indicators, applying mean imputation to one-hot encoded categorical attributes results in somewhat better performance than mode imputation.

On the basis of the experiments in this paper, we conclude that using missing-indicators to combine mean/mode imputation with the representation of missing values is a safe default approach towards missing values in classification tasks. While over- or underfitting is a concern for certain classifiers, it is a concern for these classifiers with or without missing-indicators. However, practitioners may want to omit missing-indicators when the classification algorithm to be used has a special provision for missing values, when the missingness thresholds that we determined are not met, or on the basis of specific information about the distribution of missing values in the dataset. The use of missing-indicators can also be combined with feature selection and dimensionality reduction algorithms to increase the information density of the resulting dataset.

The problem of missing data has been the subject of a rich body of theoretical literature. We hope to have contributed with this paper to the practical evaluation of some of that theory. In particular, we are happy to have identified twenty real-life datasets with missing values, and hope that in the future, more such datasets will be collected, which would allow drawing even firmer conclusions.

References

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó, 1971.

Paul David Allison. *Missing Data*. Sage Publications, Thousand Oaks, California, 2001.

Paul David Allison. Missing data. In Peter V Marsden and James D Wright (eds.), *Handbook of Survey Research*, chapter 20, pp. 631–657. Emerald Group Publishing, Bingley, England, second edition, 2010.

Andy B Anderson, Alexander Basilevsky, and Derek P J Hum. Missing data: A review of the literature. In Peter Henry Rossi, James D Wright, and Andy B Anderson (eds.), *Handbook of Survey Research*, Quantitive Studies in Social Relations, chapter 12, pp. 415–494. Academic Press, New York, 1983.

Marco Aste, Massimo Boninsegna, Antonino Freno, and Edmondo Trentin. Techniques for dealing with incomplete data: a tutorial and survey. *Pattern Analysis and Applications*, 18(1):1–29, 2015.

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

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. The Wadsworth statistics/probability series. Wadsworth, Monterey, California, 1984.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011.

Laurent Candillier and Vincent Lemaire. Design and analysis of the nomao challenge: Active learning in the real-world. In *ECML-PKDD 2012: Active Learning in Real-world Applications Workshop*, 2012.

Bojan Cestnik, Igor Kononenko, and Ivan Bratko. ASSISTANT 86: A knowledge-elicitation tool for sophisticated users. In *EWSL 87: Proceedings of the 2nd European Working Session on Learning*, pp. 31–45. Sigma Press, 1987.

Winston K Chow. A look at various estimators in logistic models in the presence of missing values. Technical Report N-1324-HEW, Rand Corporation, Santa Monica, California, 1979. URL `https://www.rand.org/pubs/notes/N1324.html`.

Jacob Cohen. Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70(6):426–443, 1968.

Jacob Cohen and Patricia Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.*, chapter 7. Missing Data, pp. 265–290. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1975.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

David Roxbee Cox. Some procedures connected with the logistic qualitative response curve. in (fn david, ed.) research papers in statistics: Essays in honour of j. neyman's 70th birthday. In Florence Nightingale David (ed.), *Research Papers in Statistics: Festschrift for J. Neyman*, pp. 55–71. John Wiley & Sons, London, 1966.

Swagatam Das, Shounak Datta, and Bidyut B Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5): 304–310, 1989.

Yufeng Ding and Jeffrey S Simonoff. An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, 11(1):131–170, 2010.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Sahibsingh A Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, 6(4):325–327, 1976.

Bradley Efron and Gail Gong. Statistical theory and the computer. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 3–7. Springer, 1981.

Emil Eirola. *Machine learning methods for incomplete data and variable selection.* PhD thesis, Aalto University, Espoo, 2014.

Matthias Elter, Rüdiger Schulz-Wendtland, and Thomas Wittenberg. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical physics*, 34(11):4164–4172, 2007.

Craig K Enders. *Applied Missing Data Analysis.* Methodology in the Social Sciences. The Guilford Press, New York, 2010.

Bob Evans and Doug Fisher. Overcoming process delays with decision tree induction. *IEEE Expert*, 9(1): 60–66, 1994.

Camila Ferreira Costa and Mario A Nascimento. IDA 2016 industrial challenge: Using machine learning for predicting failures. In *IDA 2016: Proceedings of the 15th International Symposium on Intelligent Data Analysis*, volume 9897 of *Lecture Notes in Computer Science*, pp. 381–386. Springer, 2016.

Evelyn Fix and JL Hodges, Jr. Discriminatory analysis — nonparametric discrimination: Consistency properties. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951. URL `https://apps.dtic.mil/sti/citations/ADA800276`.

Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *EuroCOLT '95: Proceedings of the Second European Conference on Computational Learning Theory*, volume 904 of *Lecture Notes in Computer Science*, pp. 23–37. Springer, 1995.

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*, volume 72 of *Intelligent Systems Reference Library*, chapter 4. Dealing with Missing Values. Springer, Cham, Zug, 2015.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1): 3–42, 2006.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS 2010: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Sergey E Golovenkin, Jonathan Bac, Alexander Chervov, Evgeny M Mirkes, Yuliya V Orlova, Emmanuel Barillot, Alexander N Gorban, and Andrei Zinovyev. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. *GigaScience*, 9(11):giaa128, 2020.

John W Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:549–576, 2009.

Jerzy W Grzymala-Busse. Knowledge acquisition under uncertainty—a rough set approach. *Journal of Intelligent and Robotic Systems*, 1(1):3–16, 1988.

Jerzy W Grzymala-Busse and Ming Hu. A comparison of several approaches to missing attribute values in data mining. In *RSCTC 2000: Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing*, volume 2005 of *Lecture Notes in Artificial Intelligence*, pp. 378–385. Springer, 2000.

H Altay Güvenir, Burak Acar, Gülşen Demiröz, and Ayhan Çekin. A supervised machine learning algorithm for arrhythmia analysis. In *Proceedings of the 24th Annual Meeting of Computers in Cardiology*, volume 24 of *Computers in Cardiology*, pp. 433–436. IEEE, 1997.

David J Hand and Robert J Till. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

Geert J M G van der Heijden, A Rogier T Donders, Theo Stijnen, and Karel G M Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology*, 59(10):1102–1109, 2006.

John D Hutcheson, Jr and James E Prather. Interpreting the effects of missing data in survey research. *Southeastern Political Review*, 9(2):129–143, 1981.

Michael P Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

Jae-On Kim and James Curry. The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2):215–240, 1977.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR 2015: 3rd International Conference on Learning Representations*, 2014. URL `https://arxiv.org/abs/1412.6980`.

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. AAAI Press, 1996.

Julián Luengo, Salvador García, and Francisco Herrera. On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32(1):77–108, 2012a.

Julián Luengo, José A Sáez, and Francisco Herrera. Missing data imputation for fuzzy rule-based classification systems. *Soft Computing*, 16(5):863–881, 2012b.

Benjamin M Marlin. *Missing Data Problems in Machine Learning*. PhD thesis, University of Toronto, 2008.

Michael McCann, Yuhua Li, Liam Maguire, and Adrian Johnston. Causality challenge: benchmarking relevant signal components for effective monitoring and process control. In *NIPS 2008: Proceedings of Workshop on Causality*, volume 6 of *Proceedings of Machine Learning Research*, pp. 277–288. JMLR Workshop and Conference Proceedings, 2008.

Mary McLeish and Matthew Cecile. Enhancing medical expert systems with knowledge obtained from statistical data. *Annals of Mathematics and Artificial Intelligence*, 2(1–4):261–276, 1990.

Ryszard S Michalski and Richard L Chilausky. Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *International Journal of Policy Analysis and Information Systems*, 4(2):125–161, 1980.

Chong Guan Ng and Muhamad Saiful Bahri Yusoff. Missing values in data analysis: Ignore or impute? *Education in Medicine Journal*, 3(1), 2011.

John G Orme and Janet Reis. Multiple regression with missing data. *Journal of Social Service Research*, 15 (1–2):61–91, 1991.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

António Pereira Barata, Frank W Takes, H Jaap van den Herik, and Cor J Veenman. Imputation methods outperform missing-indicator for data missing completely at random. In *ICDM 2019: Proceedings of the Workshops*, pp. 407–414. IEEE, 2019.

Alexandre Perez-Lebel, Gaël Varoquaux, Marine Le Morvan, Julie Josse, and Jean-Baptiste Poline. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11(1):giac013, 2022.

Therese D Pigott. A review of methods for missing data. *Educational Research and Evaluation*, 7(4):353–383, 2001.

John Ross Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.

John Ross Quinlan, Paul J Compton, K A Horn, and Leslie Lazarus. Inductive knowledge acquisition: a case study. In *Proceedings of the Second Australian Conference on Applications of Expert Systems*, pp. 157–173. Turing Institute Press, 1986.

Frank Rosenblatt. Principles of neurodynamics — perceptrons and the theory of brain mechanisms. Technical Report VG-1196-G-8, Cornell Aeronautical Laboratory, Buffalo, New York, 1961. URL `https://apps.dtic.mil/sti/citations/AD0256582`.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

L Jerlin Rubini and Perumal Eswaran. Generating comparative analysis of early stage prediction of chronic kidney disease. *International Journal of Modern Engineering Research*, 5(7):49–55, 2015.

Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J García-Laencina, Adélia Simão, and Armando Carvalho. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of biomedical informatics*, 58:49–59, 2015.

Joseph L Schafer. *Analysis of Incomplete Multivariate Data*, volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

Joseph L Schafer and John W Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

Jeffrey Curtis Schlimmer. *Concept Acquisition Through Representational Adjustment*. PhD thesis, University of California, Irvine, 1987.

Pouya Soltani Zarrin, Niels Röckendorf, and Christian Wenger. In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools. *IEEE Access*, 8:168053–168060, 2020.

Matthew Sperrin, Glen P Martin, Rose Sisk, and Niels Peek. Missing data should be handled differently for prediction than for description or causal explanation. *Journal of Clinical Epidemiology*, 125:183–187, 2020.

Stephen A Stumpf. A note on handling missing data. *Journal of Management*, 4(1):65–73, 1978.

Michael E Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

Peter Vamplew and Anthony Adams. Missing values in a backpropagation neural net. In *ACNN '92: Proceedings of the Third Australian Conference on Neural Networks*, pp. 64–66. Sydney University Electrical Engineering, 1992.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Ji Zhu, Hui Zou, Saharon Rosset, and Trevor Hastie. Multi-class AdaBoost. *Statistics and Its Interface*, 2 (3):349–360, 2009.