A Survey of Statistical Imputation Methods for Use in Machine Learning

Michael Price

Alex Xiong

Statistics 413

Rice University

December 4, 2020

Introduction

Missing data is a prevalent problem throughout data-exploration and model-building in all applications. During the process of data collection, there may be a systematic or random reason for the non-inclusion of specific data points. The naive way to account for missing/incomplete data would be to remove the data point during cleaning, as there is no knowledge as to what that data point would be precisely. However, incomplete data points may only be omitted if the data is missing at random and only a few data points are omitted. Otherwise, the would be bias in the resulting sampling of data as well as a heavy reduction of data, respectively. To account for both of these issues, there have been many imputations methods proposed in literature that can be used to infer what the missing data points would be from the existing data. These methods will be further explained in the Explanation of Imputation Methods section.

The types of missing data can be split into three categories: Missing completely at random (MCAR), Missing at random (MAR), and Not missing at random (NMAR) (Mack 2018). Data missing completely at random denotes that the missing data is independent to the non-missing data. Therefore, the non-missing data is still a representative sample of the population. Data missing at random denotes that the incomplete data is dependent on some measured predictors. This bias can be corrected if properly accounted for. The final category, data not missing at random, means that data has been collected with undetermined factors affecting the process, resulting in missing data. This may result in non-correctable biased sampling.

For this analysis, we investigated the effect of different imputation techniques on multiple datasets with missing completely at random data. Moreover, to present novelty in our approach, we implemented an ensemble imputation technique that combined the results multiple different imputation techniques. The intention of this ensemble is to decrease variability within imputation calculations while only training on a single dataset, rather than relying on multiple models from different training sets.

Explanation of Imputation Methods

MICE

MICE (Multiple Imputation by Chained Equations) belongs to a family of imputation techniques known as multiple imputation methods. All algorithms of this type follow a similar strategy, which was first proposed by Donald Rubin (1978). In general, instead of

performing analysis on one imputed dataset, multiple imputed datasets are created. A model is trained on each of those imputed datasets, and the results are aggregated (Azure et al. 2011). To ensure that each imputed dataset is unique, the imputed values are drawn from some probability distribution. In theory, by utilizing multiple different datasets for training, the aggregated model will have low variance, similar to how bagging can help reduce variance in machine learning models.

In summary, MICE is a Bayesian method which implements a Gibbs sampler to draw possible imputed values. A basic explanation of the algorithm is provided below:

1. First, the missing values, y_{ij} are imputed with an initial guess, usually a sample mean if the variable is numerical and a mode if the variable is categorical.

2. For each predictor X_j, the posterior distribution of a Bayesian Generalized Linear Model is computed, $p(\theta|X_j, X_{-j})$, where θ is a vector of parameters for the model, and with X_j as the response variable and all other variables X_{-j} as the regressors. Next, m values are sampled from the posterior distribution (θ_{ij}), where m is the number of missing values (y_{ij}) in X_j.

3. Finally, new values of y_{ij} are sampled from the likelihood, $p(y_{ij}|X_j, X_{-j}, \theta_{ij})$. This sampled value will replace the previous imputed value.

4. This process is repeated for each predictor multiple times. Ideally, the sampler will converge to the "true" values of y_{ij}, thus resulting in an accurate estimation of the missing values.

5. The process above is repeated multiple times, therefore resulting in multiple imputed datasets.

In practice, unlike many other Gibbs Samplers, only about 10-20 iterations are needed for convergence (van Buuren and Groothuis-Oudshoorn 2011), thereby reducing computation time. The number of datasets outputted is entirely up to the user, in which variables like size of dataset should be considered.

KNN Imputation

A potential downside of MICE is that it is a computationally expensive approach, given the multiple imputations and the fact that a model must be trained on each of the outputted datasets. Therefore, especially in the case when the dataset is large, an algorithm that outputs only one imputed dataset is desirable. KNN Imputation fits this criteria, and utilizes

the popular and simple KNN algorithm. (Kowarik and Templ 2016). However, given that multiple different types of variables (continuous, binary, ordinal, etc.) may exist in the dataset, KNN Imputation utilizes the Gower distance (Gower 1973) instead of the traditional Euclidean distance. The Gower distance d_{ij} between observations x_i and x_j is defined as

$$d_{ij} = \frac{\sum_{\ell=1}^{p} w_\ell \delta_{ij\ell}}{\sum_{\ell=1}^{p} w_\ell}$$

where $\delta_{ij\ell}$ changes based on the type of variable. For example, for binary variables, it is defined to be

$$\delta_{ij\ell} = \begin{cases} 1 & x_{i\ell} = x_{j\ell} \\ 0 & x_{i\ell} \neq x_{j\ell} \end{cases}$$

In general, no matter the variable, the $\delta_{ij\ell}$ is in between 0 and 1, thus ensuring that each variable in the dataset is on a similar scale. w_ℓ is a user defined weight for variable ℓ, where if a user deems a variable "more important", w_ℓ will have a relatively large value compared to other weights. Finally, to predict the missing value, the k nearest observations according to the Gower distance are considered. If the variable is numerical, the median of the neighbors is outputted, and if the variable is categorical, the algorithm outputs a majority vote.

MissForest

Nonparametric Missing Value Imputation Using Random Forest, otherwise known as MissForest, is an implementation of the random forest algorithm applied specifically to imputation. While a majority of imputation methods cannot handle mixed-type data, MissForest is equipped to handle continuous and/or categorical data. Furthermore, due to the usage of random forests, MissForest can account for interactive and non-linear effects (Stekhoven 2011). MissForest works by building a random forest for each variable. The missing values are predicted by the random forest model iteratively. Moreover, the out of bag (OOB) error rates for random forests can be calculated without a test set, ensuring that the quality of imputation can be tested without reserving testing data or performing cross validation. A basic explanation of the algorithm is as follows:

1. Make an initial guess for the missing values.

2. Sort the indices of X, a $n \times p$ matrix of predictors, by increasing amount of missing values.

3. Iterate until the stopping criterion is met, for every predictor, iterate and fit a random forest between the predictors and the response on the known data.

4. Predict the missing data based on the random forest model and update the imputed matrix.

5. Once the stopping criterion is met, return the imputed matrix.

Ensemble Approach

In "big data" situations, multiple imputation methods may be unfeasible, given the need to train models on multiple datasets. Therefore, standard procedures like hyperparameter tuning and cross-validation will require an immense amount of computation. Thus, MissForest and KNN are attractive options for imputation because they output only one dataset. However, because only one dataset is outputted, these methods may suffer from high variability.

Therefore, it would convenient for a data analyst faced with missing data to have access to an imputation procedure that outputs one dataset, but also does not suffer from high variability. To achieve this, we propose an ensemble method that takes into account the results from the three imputation algorithms listed above. Essentially, MICE, and multiple versions of KNN and missForest with varying hyperparameters are run on the same dataset. Then, each imputed values for each outputted dataset will be aggregated in some way, and that aggregation will be value imputed into the dataset. For numerical features, we propose using a sample median to mitigate the effect of possible outliers, and for categorical variables, we propose taking a majority vote. If there is a tie, then one answer will be chosen randomly. With this technique, we hope to achieve lower variance than other approaches that output one dataset, as more models are taken into consideration.

Experimental Setup

Imputation Experiment Overview

For this experiment, we are primarily interested in how each imputation method affects the predictive power of a variety of machine learning models. In addition, we also are interested in how varying the amount of data missing in the dataset affects the performance of the imputation algorithms. For this experiment, we implemented the following procedure in R on 3 complete datasets, varying the imputation procedure and percent of data removed (1%, 5%, and 10%):

1. Split the data into training and testing sets. The sets are chosen randomly.

2. Randomly remove $r\%$ of the data points from the training set, where r is in between 0 and 100.

3. Impute the missing values with some imputation algorithm.

4. Train a variety of models with the imputed dataset.

5. Test the accuracy of the trained models with the testing set. Record the error rate for classification tasks and the mean squared error for regression.

This process is repeated multiple times (100 in our experiment) in order to obtain a sampling distribution of errors for each model. We will then compare the distribution of these errors among the different imputation methods, mostly focusing on the mean and the standard deviation, as these will be our measures of "bias" and "variance". Ideally, these imputation techniques will yield both low testing error and error variance.

For this analysis, the `mice` package was used for MICE, the `VIM` package for KNN, and the `missForest` package for MissForest. In addition, each imputation method discussed earlier has a variety of hyperparameters, example being the number of trees from MissForest. Although metrics exist to asses the quality of the imputation, like testing if the distribution of the imputed values match the distribution of the non-missing values, the default hyperparameters within each R function were mostly used throughout this analysis. The full list of hyperparameters is listed below:

Imputation Method	Hyperparameters
MICE	*Iterations* = 10, *Imputed Datasets* = 5
KNN	$k = 5$
MissForest	*Variables Considered per Tree* = $\sqrt{Number\ of\ Predictors}$, *Number* of Trees = 100
Ensemble	*Number of Datasets* = 9, 5 from MICE, 2 from KNN, and 2 from MissForest Same Hyperparamters from above, including $k = 3$ and *Number of Trees* = 150

Table 1: Imputation Hyperparameters

In addition, we included another imputation method in the experiment as a baseline, which we named "Simple". The algorithm simply imputes the sample mean of the non-missing values if the variable is numeric and the sample mode of the non-missing values if the variable is categorical.

Data Used

We implemented the procedure outline above on 3 datasets, with 2 being a classification task and 1 being a regression task. Each of them are described below:

1. Wage - This dataset, containing 9 variables and 3000 observations, is available in the `ISLR` package and documents the wage and personal information of various working men across the Middle Atlantic Region of the United States. We will attempt to predict the log of the worker's wage given the other variables.

2. Diabetes - The Diabetes dataset, containing 9 variables and 768 observations, comes from a 1988 study documenting whether various women of Pima Indian decent had diabetes. We will attempt to predict if the woman has diabetes given factors like her age, number of children, and BMI.

3. Steel Plates - This dataset containing 28 variables and 1941 observations, contains information of various faults in order to classify them into two categories - common or other. Examples of variables included in the dataset are include orientation, size, and location.

Before further testing, LASSO Regression was performed on each of the datasets to remove the insignificant variables.

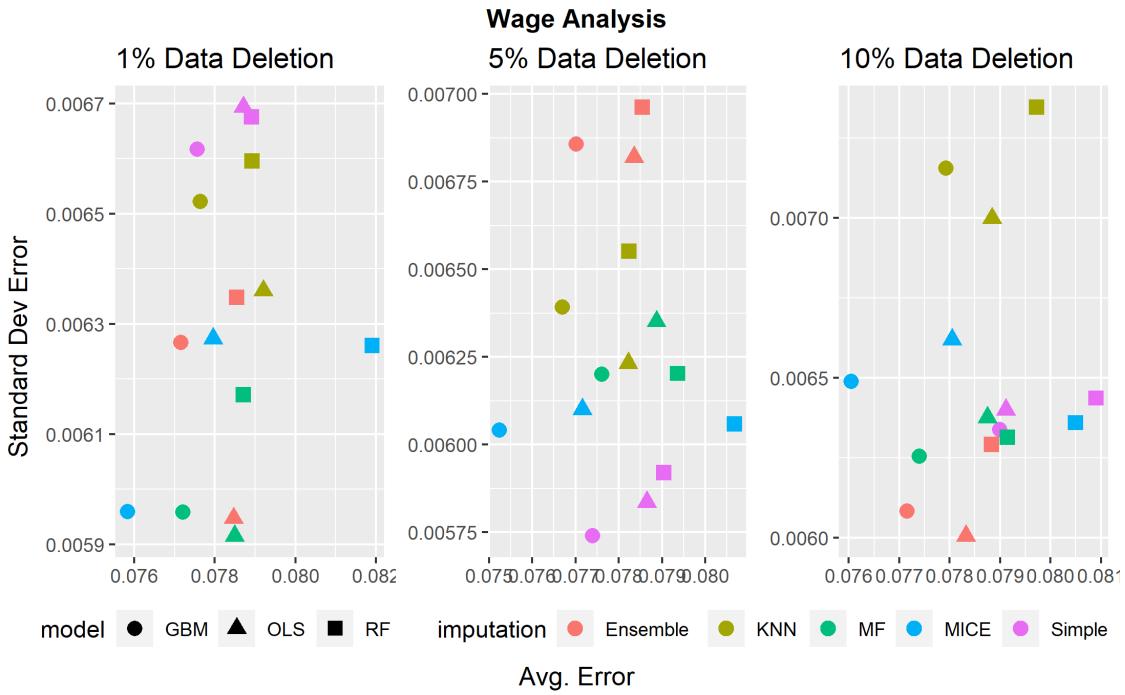
Models Used

For regression, we implemented 3 models - Ordinary Least Squares Regression, Random Forest and Gradient Boosting. For classification, we implemented 5 different models - Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine and KNN. We performed standard grid search and cross validation on the entire complete dataset to identify the optimal set of hyperparameters for each of the models. The models' hyperparameters remained constant throughout the entire testing procedure.

Results

For each of the datasets, we plotted the average error vs. the standard deviation of error from each imputation, model pair. Therefore, the best imputation model pairs are located in the bottom left corner of the result plot, while the worst performing pairs are located on the top right. The results are stratified by the percentage of data deletion.

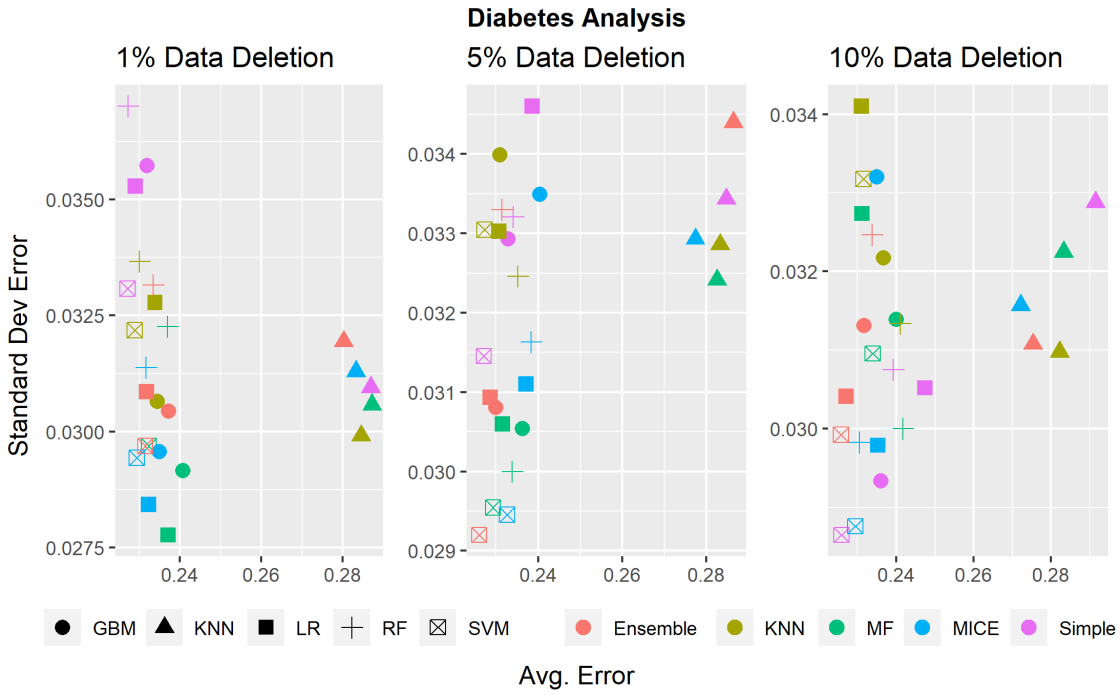
Wage Results



In the Wage plot, we observe that MissForest is the best overall balance of mean and the standard deviation of error for 1% data deletion, given it achieves low variance for each of the models. For 5% data deletion, simple imputation achieved the lowest variance, while average error was similar for all the imputation methods. For 10% data deletion, instead, the ensemble imputation had the lowest variance, while no single imputation method achieved noticeably smaller error.

We can also observe that our novel ensemble method is competitive with MICE in both the 1% and 10% data deletion cases. In the scope of the Wage dataset, this demonstrates that we do not necessarily need to train models on multiple datasets to reduce variance for imputation. However, given the poor performance of the ensemble method at the 5% data deletion, we can not make a definitive statement on whether it is always a better or worse solution compared to other imputation methods.

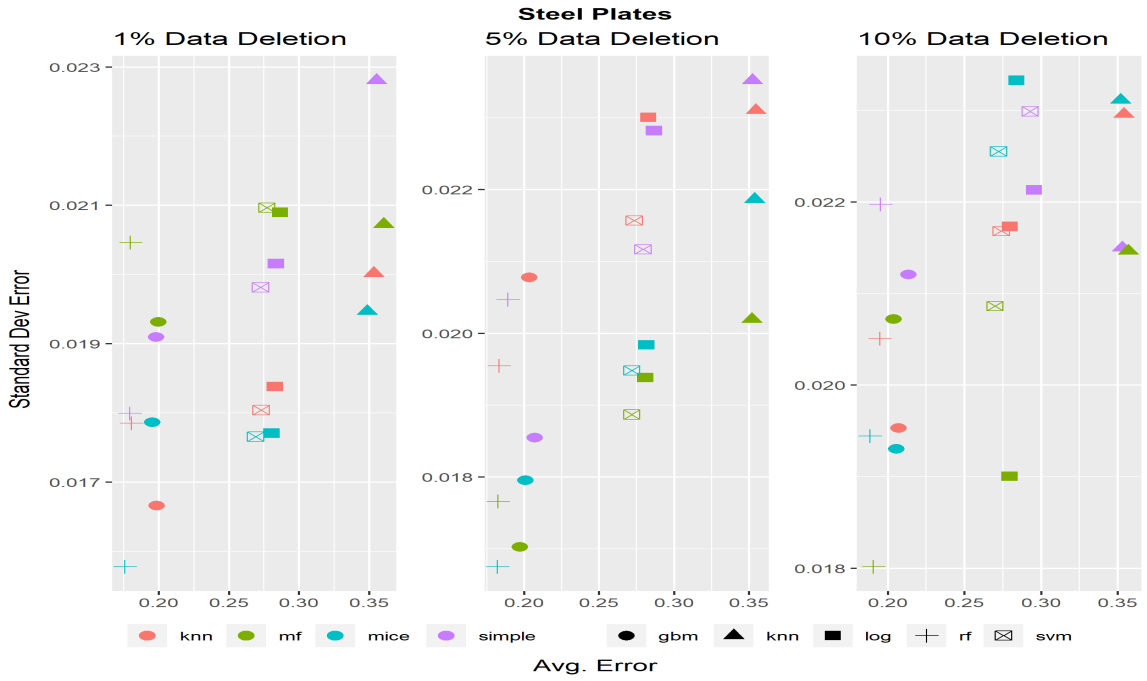
Diabetes Results



In the Diabetes plot, we observe that MissForest is the best overall balance of mean and the standard deviation of error for 1% and 5% data deletion, given the relatively variance. For 10% data deletion, no imputation technique achieves separation from the others over all models, thus no definite conclusion can me made on which one performed best.

We can also observe that our novel ensemble method is competitive with MICE in both the 1%, 5%, and 10% data deletion cases, given the similar average error and standard deviation of error. Thus, like Wage, multiple imputation methods are not strictly needed to improve performance.

Steel Plates Results



In the Steel Plates plot, we observe that MICE is the best overall balance of mean and the standard deviation of error for 1% data deletion given the low variability. For similar reasons, for 5% and 10% data deletion, we can conclude MissForest performed the best. Due to the computational demands of fitting the ensemble method multiple times on the dataset, we were unable to extract any Ensemble results from the Steel Plates dataset. In conclusion, given that the results vary greatly among the three datasets, we can definitively conclude that no single imputation method consistently outperforms others.

Conclusion

In this analysis, we examined four conventional data imputation techniques as well as a novel ensemble imputation method. While no algorithm consistently performed the best among the datasets, we were still able to make basic distinctions within each dataset between different the different imputation methods. In the future, it would be beneficial to examine more data deletion percentages to gain a wider perspective on how the data deletion split affects the standard deviation of error with respect to average error. Furthermore, we could also examine different types of datasets, focusing on numerical, categorical, and mixed-data datasets to distinguish any relationship between the performance of our ensemble methods and other imputation techniques. Another possible step would also be to quantify the optimal imputation technique across all models.

The application of the ensemble imputation method on the Wage and Diabetes dataset showed promising results, especially on datasets with numerous types of variable. Thus, we can conclude that multiple imputation methods like MICE are not surefire ways to reduce variance. However, the Ensemble method is still relatively computationally expensive and its performance is comparable to MissForest and KNN. Thus, it is certainly not appropriate for all situations. The future paths mentioned above should be pursued as well with respect to ensemble imputation to investigate its viability further.

Although the ensemble imputation method shows promise against methods such as MICE and MissForest, no one method proved to outperform the others in all cases. Therefore, when performing real-world analysis, we recommend comparing multiple different imputation methods to gain a more complete view on how each imputation method affects model training and evaluation.

References

Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. https://doi.org/10.1002/mpr.329

Gower, J. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4), 857-871. doi:10.2307/2528823

Kowarik, A., Templ, M. (2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7), 1 - 16. doi:http://dx.doi.org/10.18637/jss.v074.i07

Rubin, D.B. (1978) Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association, 20-28.

Mack C, Su Z, Westreich D. Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018 Feb. Types of Missing Data. Available from: https://www.ncbi.nlm.nih.gov/books/NBK493614/

Stekhoven, D, J., Bühlmann, P., MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics, Volume 28, Issue 1, 1 January 2012, Pages 112–118, https://doi.org/10.1093/bioinformatics/btr597

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1 - 67.
doi:http://dx.doi.org/10.18637/jss.v045.i03

Datasets

Diabetes (https://www.openml.org/d/37) and Steel Plates (http://archive.ics.uci.edu/ml/datasets/steel+plates+faults) Datasets provided by UCI Machine Learning Depositiory

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

Irvine, CA: University of California, School of Information and Computer Science.

Wage Dataset provided by Introduction to Statistical Learning with R textbook

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York