Model based Data Imputation

Vittanala Sai Bhushan
Student
Department of CSE
GVPCE

P. Krishna Subba Rao
Professor
Department of CSE
GVPCE

ABSTRACT
Missing or incomplete data is a significant problem in all types of statistical analyses. In this project, multiple imputations using chained equation (MICE) is modified to work with various regression algorithms such as linear regression algorithm. The modified MICE algorithm then will be compared using accuracy on three different datasets.

General Terms
Imputation, Machine Learning, Python.

Keywords
Data Imputation, MICE, Machine learning, Multiple Imputation, Random Forest.

1. INTRODUCTION
Single imputation.

The missing values are imputed only once based on rules provides estimates based on observed score of the variables for which the data is missing.

Types of techniques present in single value imputation.

Replace with mean or median:-

The mean is calculated on non missing values and the result is filled near the missing data.

Replace with most frequent value:-

The most frequent values present in the column are used to fill the missing data.

Hot-deck imputation:-

To impute any random values from any similar records in place of missing values in the dataset. Hot-deck imputation is last observation carried forward. To helps sorting the dataset based on the number of variables.

Cold-deck imputation:-

To impute values that are taken from another similar dataset in place of missing values in the present dataset.

Multiple imputation.

This method is used to deal with the problem of increased noise due to imputation.

It follows three steps.

Step 1:

Imputation:-

The missing values are imputed multiple times forming multiple datasets.

Step 2:

Analysis:-

The multiple datasets are analyzed such that they generate multiple sets of outcomes.

Step 3:

Polling:-

All the results generated are combined into single multiple imputationresults.

Multiple imputation methods

Multivariate Imputation by Chained Equation(MICE).

In this imputation, the missing values are imputed using mean in the beginning. In the next step, remove the imputed value from the target variable and then apply linear regression on the other fully filled variables such that the missing data row becomes the test data. After predicting the value fill the missing value and perform this process on other variables with missing data until all the values are filled. Next, subtract the imputed dataset values with the actual values. This process continues until the resultant dataset values will be nearly equal tozeroes.

2. BACKGROUND AND RELATED WORK

➢ **In 2008Markov chain Monte Carlo Multiple Imputation**

- A large survey dataset is taken with missing values.
- In this dataset, the missing values are present with complex pattern.
- The MCMC technique is used to impute them as it has good convergenceproperties.
- The results stated that imputation with MCMC gave accuratevalues.

➢ **In 2010 Mean ValueImputation.**

- An yearly income dataset of 4162 people is taken which has missing values in yearly income.
- Two simple imputation techniques and one multiple imputation technique is used for imputation andcomparison.
- Here, the comparison is done based on AUC curve values.
- The multiple imputation technique gave better AUC value of 0.73.
- This also stated the association between annual income and Self Reported Healthstatus.

➢ **In 2011 --- Multiple Imputation by Chained Equations.**

- This paper was written to help psychiatric researchers for imputing complex patterned missing values.
- This research helped them to understand the

principle behind MICE imputation and the software available to apply it on dataset.

- **In 2014 --- MultipleImputation.**

- This research took clinical data of 20 TB patients. In this data, few values present in the 8 parameters weredeleted.
- These values are imputed using single and multiple imputationtechniques.
- Where, Multiple Imputation gave better values and outperformed othertechniques.

- **In 2019 --- Review on Types of Missing Values and Multiple ImputationTechniques**

- This paper tells us about various types of missing values such as MAR,MNAR.
- Then it also states about the views of various other researchers on missingvalues.
- Then it introduces various multiple imputation techniques that can be used forimputation.

- **In 2020 --- Different tensor model techniques**

- In this research, four tensor completion models including Smooth PARAFAC Decomposition based Completion (SPC), CP Decomposition-based (CP- WOPT) Completion, Tucker Decomposition-based Completion (TDI), and High-accuracyLow-rankTensor Completion (HaLRTC) are used to impute the missing data.
- Then the prediction models Support Vector Regression (SVR), K-nearest Neighbor (KNN), Gradient Boost Regression Tree (GBRT), and Long Short-term Memory (LSTM) are tested on the imputed data.
- The results showed that increasing the accuracy of missing data gives betterresults.

2.1 MACHINE LEARNING MODELS FOR ANALYSIS

2.1.1 RANDOM FOREST

Random forest is a supervised machine learning algorithm used for both classification and regression. It takes the complete dataset as input and divides it into various subsets randomly. It builds decision trees from these subsets. The prediction of each decision tree is considered and the class with highest prediction is given for the testdata.[12].

3. MATERIAL AND METHODS

3.1 Dataset

There are three different datasets of house data, horse data and travel-time data. The algorithms impute the data and evaluates based on mean absolute error.

3.2 Model Overview

Figure 1. describes the proposed model that consists of Pre-processing, Classification and Evaluation phases which are explained below,

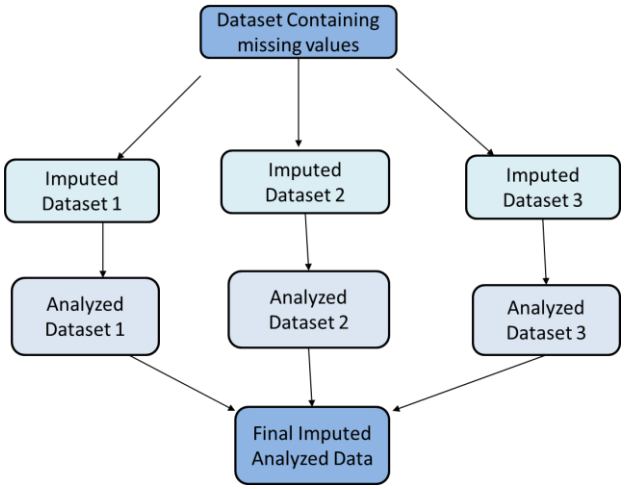

Fig 1: System Architecture

3.2.1 Pre-Processing

In the proposed model, the pre-processing is done by filling the missing values present in the dataset. Each dataset contains missing values in atleast one attribute. For example, Housing Dataset has missing values in the attribute LotFrontage. Similarly, in other two datasets there are few missing values present in theirattributes.

3.2.2 Classification Techniques

In this project, The classifier models constructed are Random forest, regressor for imputing the missing values. These classifier has been discussed in the previous section, and the evaluation of their performance is carried out in the next section.

4. PERFORMANCE METRICS

The performance of the algorithm is evaluated based on the mean absolute error.

4.1 MEAN ABSOLUTE ERROR

Mean Absolute Error is the model evaluation metric used in regression models. The mean absolute error of the model for the test set is the average of the absolute values of the individual prediction errors across all the instances in the test set.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

5. RESULTS AND DISCUSSION

In this section, we have analysed the parameter (Mean absolute Error) on Random Forest classifier based on three datasets.

Table. Comparison Of Machine Learning Algorithm With Respect To The Given Problem Statement on threedatasets

Housing Dataset

Drop Missing Values	Single Imputation	Multiple Imputation
0.083	0.0821	0.081

Horse Dataset

Drop Missing Values	Single Imputation	Multiple Imputation
0.296	0.277	0.271

Travel-time Dataset

Drop Missing Values	Single Imputation	Multiple Imputation
0.136	0.135	0.127

The proposed model uses a machine learning technique that were set to achieve better imputation of values. As per the given results, it can be inferred that the multiple imputation gave the least mean absolute error(0.081) and single imputation gave second least mean absolute error(0.082) on the given dataset. Whereas in Horse dataset, as per the results, it can be inferred that the multiple imputation gave the least mean absolute error(0.271) and single imputation gave second least mean absoluteerror(0.272).

It can also be inferred that, the multiple imputation gave the least mean absolute error(0.127) and single imputation gave second least mean absolute error(0.135) while dropping of missing values gave highest value(0.136).

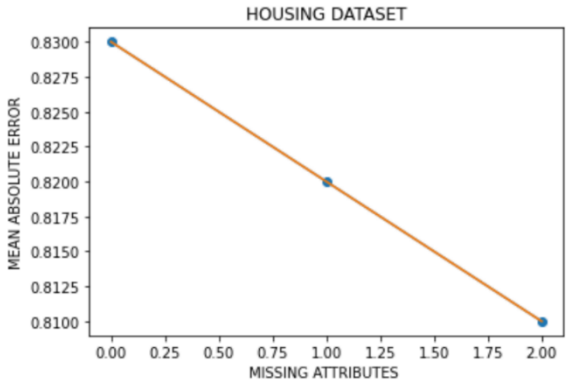

Figure. 2 Line chart of accuracy with respect to mean absolute error on House Dataset

From the above figure 2, it can be inferred that, Multiple Imputation gave least mean absolute error 0.9277 and single imputation gave mean absolute error 0.9274 respectively compared to dropping of missing values.

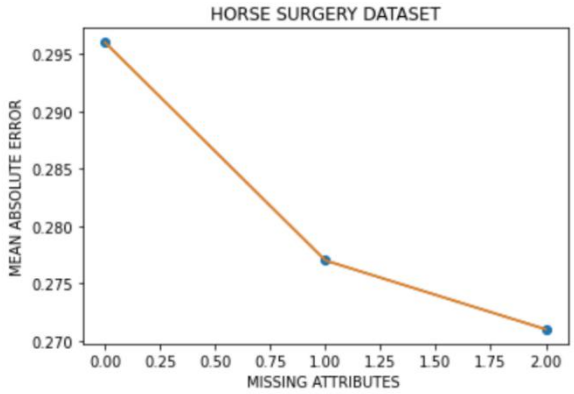

Figure. 3 Line chart of mean absolute error with respect to horse dataset

From the above figure 3, it can be inferred that, multiple imputation gave least mean absolute error of 0.271 while single imputation gave least mean absolute error of 0.277 respectively when compared to dropping of missing values.

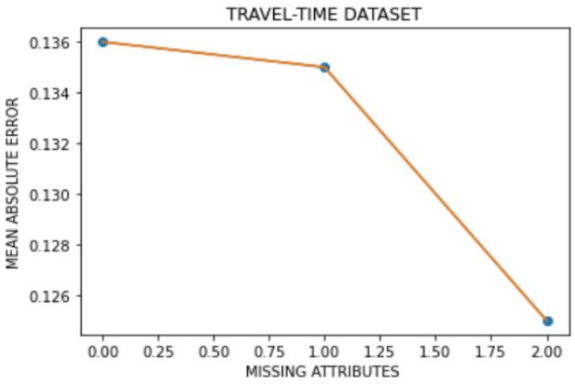

Figure. 4 Line chart of mean absolute error based on travel-time dataset.

From the above figure 4, it can be inferred that, multiple imputation gave better mean absolute error of 0.125 while single imputation gave 0.135 when compared to dropping of missing values.

6. CONCLUSION AND FUTURE ENHANCEMENT

Based on the above results, in all the three datasets, the mean absolute error is least while applying multiple imputation rather than while applying single imputation and by dropping missing values.

Hence it can be concluded that multiple imputation with MICE can be used efficiently for imputing the missing values in these threedatasets.

The other multiple imputation techniques such as MCMC, MIMCA etc can also be applied on these datasets and their efficiency can be tested using any machine learning technique.

7. REFERENCES

[1] D. Schunk, "A Markov chain Monte Carlo algorithm for multiple imputation in large surveys", AStA Advances in Statistical Analysis, vol. 92, no. 1, pp. 101-114,2008.

[2] A.Ryder et al., "The Advantage of Imputation of Missing Income Da ta to Evaluate the Association Between Income and Self-Reported Health Status (SRH) in a Mexican American Cohort Study", Journal ofImmigrant

[3] M. Azur, E. Stuart, C. Frangakis and P. Leaf, "Multiple imputation by chained equations: what is it and how does it work?", International Journal of Methods in Psychiatric Research, vol. 20, no. 1, pp. 40-49,2011.

[4] C. Padgett, C. Skilbeck and M. Summers, "Missing Data: The Importance and Impact of Missing Data from Clinical Research", Brain Impairment, vol. 15, no. 1, pp. 1-9,2014.

[5] J. Pang, Y. Gu, J. Xu, Y. Bao, and G. Yu, "Efficient graph similarity join with scalable prefix-filtering using mapreduce", In Web-Age Information Management, Springer, pages 415 - 418,2014.

[6] X. Zhao, C. Xiao, W. Zhang, X. Lin, and J. Tang, "Improving Performance of Graph Similarity Joins Using

Selected Substructures", Springer International Publishing, Cham, pages 156 -172,2014.

[7] Audigier, F. Husson and J. Josse, "MIMCA: multiple imputation for categorical variables with multiple correspondence analysis", Statistics and Computing, vol. 27, no. 2, pp. 501-518,2016.

[8] J. Jakobsen, C. Gluud, J. Wetterslev and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts", BMC Medical Research Methodology, vol. 17, no. 1,2017.

[9] Jerez, J., Molina, I., García-Laencina, P., Alba, E., Ribelles, N., Martín, M. and Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2),pp.105-115.

[10] MANSOURIAN, M. and AFSHARI SAFAVI, A., 2017. Handling Missing Data in Questionnaire-Based Studies: A Comparison Between Simple and Imputation Techniques. TurkiyeKlinikleri Journal ofBiostatistics.

[11] Khan, S. and Hoque, A., 2020. SICE: an improved missing data imputation technique. Journal of Big Data, 7(1).

[12] "Imputation(statistics)",En.wikipedia.org,2020. [Online].Available:https://en.wikipedia.org/wiki/Imputation_(statistics). [Accessed: 03- Sep-2020].

[13] Numpyninja.com,2021.[Online].Available: https://www.numpyninja.com/post/mice-algorithm-to-impute-missing-values-in-a-dataset. [Accessed: 30- Mar-2021].