www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1076-1080

e-ISSN : 2583-1062
Impact Factor : 7.001

SECONDARY COMPARATIVE ANALYSIS OF AI IMPUTATION TECHNIQUES FOR MISSING VALUES IN LOAN DEFAULT PREDICTION WITH MCAR, MAR, AND MNAR PATTERNS

Prachi Purohit[1]

[1]Kishinchand Chellaram College, India.

DOI: https://www.doi.org/10.58257/IJPREMS36880

ABSTRACT

This study is done to compare some of the AI (Artificial Intelligence) techniques that have recently become popular in the process of missing value imputation. The comparison is performed on various important types of datasets (MCAR, MAR, and MNAR). Missing data is a critical problem in data analysis which if left unnoticed can cause biased predictions and conclusions. It is because of this, it is important to use the right method of imputation of the missing values. While, traditional methods like removing missing values, replacing them with central tendency (like mean, median and mode) have been used for most years, they can miss the sophistication needed for complex and real-time data. This study focuses on comparing the basic four AI imputation techniques (K-Nearest Neighbors, Random Forest Imputation, Multiple Imputation by Chained Equations, and Autoencoders) on three types of most common missing value datasets. These algorithms are then evaluated and analysed based on time taken for execution and caused errors using RMSE (Root Mean Squared Error). The results suggest that while Autoencoders may be computationally intensive, whereas simpler techniques like KNN offer efficient, moderate accuracy for certain data types.

1. INTRODUCTION

Background:

There are three main categories of missing data based on statistical literature, which are MCAR (Missing Completely at Random) – where the missingness is unrelated to any observed or unobserved variables; MAR (Missing at Random) – missingness is due to the observed variables but not the missing values themselves; and MNAR (Missing Not at Random) – missingness depends on the unobserved data itself.

Problem Statement:

In the era of big data, where data is being collected left and right, it is important to recognize the critical challenge that missingess of data can possess. Missing data can be caused due to various reasons, some basic ones include human entry error, sensor errors, respondent omissions in surveys etc. The traditional approaches like mean imputation can often oversimplify the issue and not adhere to the complexity of real-world data. Identifying what type of imputation technique work well with what kind of missing data is essential for proper analysis of the data.

2. RESEARCH OBJECTIVE

Therefore, this study aims to help identify and pose a second opinion on the effect of AI imputation techniques on the different types of missing data mentioned in earlier section on a Loan Default Prediction dataset. The results should and are intended to guide researchers in selecting the right type of imputation technique based on the kind of dataset they are working with.

Scope of Study:

Scope of this study involves analysis of missing value imputation techniques on a financial dataset with three scenarios of missing values. Research contributes to understanding the strengths and limitations of AI imputation approaches by analysing the time efficiency and accuracy of said techniques.

3. LITERATURE REVIEW

Missing data in fields like healthcare, finances, social sciences, engineering etc., have led to many biased analysis and results. Traditional methods have, thus, been used extensively to manage these missing data. However, with the advancement of AI over the few years recently and its application in different fields have led to an evolution in imputation techniques as well. Recent research has highlighted the effective impact made by the AI imputation techniques over the traditional ones.

KNN or K-Nearest Neighbors has become popular due to its simplicity and its capability in capturing the local data structures. KNN uses nearby data to impute data which results in the method being effective in datasets with relatively less dimensions. However, studies indicate that KNN's accuracy diminishes as data dimensionality increases, and it may be less effective when dealing with large volumes of missing data (Jerez et al., 2010). KNN technique finds the

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 1076-1080
e-ISSN : 2583-1062
Impact Factor : 7.001

data points that are closest to the missing value by calculating the distance using formulas like Euclidean, Manhattan. The missing value is then replaced with the average of these nearest data points. Mathematically,

$$\frac{\text{Value 1} \times \text{Weight 1} + \text{Value 2} \times \text{Weight 2} + \cdots + \text{Value n} \times \text{Weight n}}{\text{n}}$$

Where, n is the number of closest data points

Random Forest techniques have demonstrated the effectiveness and robustness in data imputation but require significantly high computational powers when dealing with the large datasets and complex models, as indicated in a research done by Stekhoven and Bühlmann (2012). It models relationships between features, filling in missing values based on predictions from multiple decision trees by leveraging ensemble learning.

MICE, or Multiple Imputation by Chained Equations, is a machine learning technique that creates multiple prediction of each missing value based on the observed data, allowing a various number of values than a single imputation. MICE is effective and versatile especially in scenarios for datasets with MAR and MCAR data. This was demonstrated very well in a research done by van Buuren (2018).

With the help of deep learning models, Autoencoders was introduced as a novel approach to the imputation of missing values. It is designed using neural networks which encodes the input data into lower dimensional space and then decode it back to its original form allowing it to learn intricate patterns in the data. A study done by Gondara and Wang (2018) found that Autoencoders performed noticeably better than the traditional methods for MAR and MNAR data in terms of accuracy. Although, this imputation technique requires a very high computational power.

With this, it is clear that not one technique is superior to others, as the strengths and limitations of each technique differ from one another and work better in a particular situation and cause biased results in another. Instead, the effectiveness of an imputation technique heavily depend on the missing mechanism, dataset characteristics and computational resources.

4. METHODOLOGY

Dataset:

The dataset used for the study is a secondary dataset, named "Default_Fin.csv" which is sourced from Kaggle and comprises financial and employment features like ("Employed, Bank Balance", "Annual Salary", and "Defaulted?"). This dataset represent a real world case scenario where the dataset can frequently encounter missing data that affects the analysis and prediction. So, three types of dataset is then evolved from this original dataset to simulate the MCAR, MAR and MNAR categories of missing data.

1. "Default_Fin_MCAR.csv": Missing data is introduced randomly which are independent of other variables or the missing values itself.
2. "Default_Fin_MAR.csv": Induced missing values based on the observed data, making it dependent on certain variables. For example, the missingness of bank balance values is based on the employment status, as the employed individuals are more likely to report their bank balance than an unemployed individual.
3. "Default_Fin_MNAR.csv": Missing values are missing because of some unobserved data, which means the data is missing because of the missing values themselves. This type of missingness is comparatively more difficult to simulate because it requires accurate reconstruction of unobserved patterns.

Imputation Techniques:

The imputation techniques to be compared are chosen based on their different approach and applicability:

1. K-Nearest Neighbors: It is a supervised machine learning technique which imputes missing data based on an average or majority of the neighbors' values.
2. Random Forest: This approach uses decision trees to predict the missing values. Each tree in this forest models the data's relationship between the features allowing it to be robust with complex data structures.
3. MICE: It is an iterative multiple imputation technique that cycles through the variables and fills the missing values. Each missing value is imputed multiple times to account for uncertainty.

Evaluation Metrics:

To assess the performance, mainly two metrics are used for analysis and comparison of techniques:

1. Root Mean Square Error (RMSE): It measures the square root of average squared differences between actual and imputed values. Lower RMSE value indicates more accuracy.
2. Time efficiency: To measure the speed and efficiency, time taken for execution of techniques is recorded. It also helps understand the computational costs, particularly for large dataset with high dimensionalities.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 1076-1080
e-ISSN : 2583-1062
Impact Factor : 7.001

5. RESULTS AND DISCUSSION

Overview:

The imputation techniques were applied to the three categories of missing data and metrics were used to evaluate the performance of each technique in each scenario of missingness. The results were then visualized using grouped bar plots for comparison of accuracy and time spent.

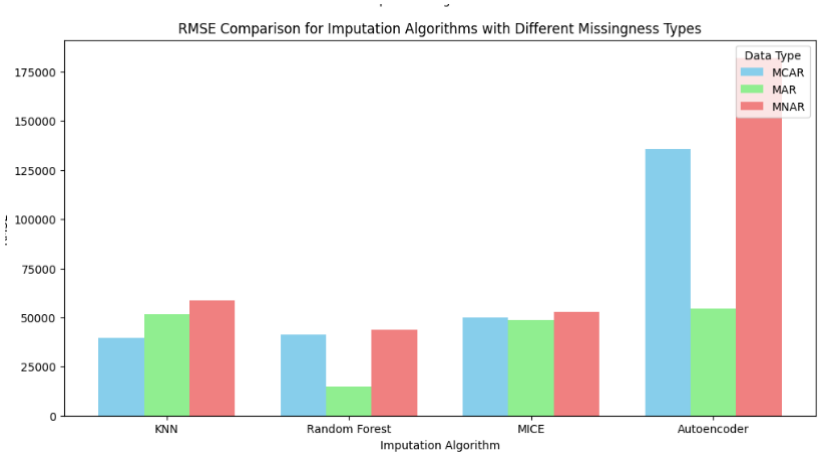

fig. 1

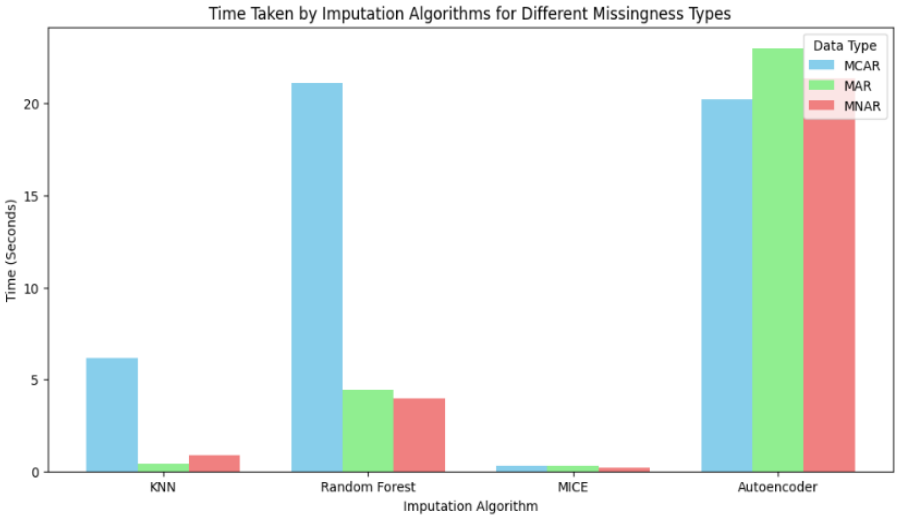

fig. 2

Performance on MCAR Dataset:

Random Forest and KNN both performed well in the MCAR scenario, with Random Forest having the lowest RMSE (41,176.36) and KNN coming in second (39,662.05). Autoencoders had the highest RMSE (135,636.05) for MCAR, according to the heatmap, which suggests they might not be the best choice for data with missing dependence patterns. The time comparison bar plot shows that MICE balanced accuracy and computing efficiency with a reasonable RMSE (50,103.82).

Performance on MAR Dataset:

When compared to other approaches, Random Forest produced the best RMSE performance (14,789.42) for the MAR data. This low error indicates that Random Forest handles observed-variable-dependent missingness quite well. Despite requiring greater calculation time, autoencoders also performed well, with a respectably low RMSE (54,554.19). Despite having greater RMSEs (51,801.16 and 48,633.66, respectively), KNN and MICE are still effective options when speed is of the essence.

Performance on MNAR Dataset:

The greater RMSEs for all approaches indicate that the MNAR dataset presented the biggest problem. The capacity of autoencoders to recreate patterns from unseen variables was reinforced by their lowest RMSE (182,112.41), albeit at a considerable computational expense. With intermediate accuracy (43,597.23 and 52,916.07, respectively), Random Forest and MICE demonstrated their limits in identifying unseen relationships. Due to its oversimplified distance-based methodology, KNN has the greatest RMSE (58,461.25), demonstrating its ineffectiveness un MNAR settings.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1076-1080
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

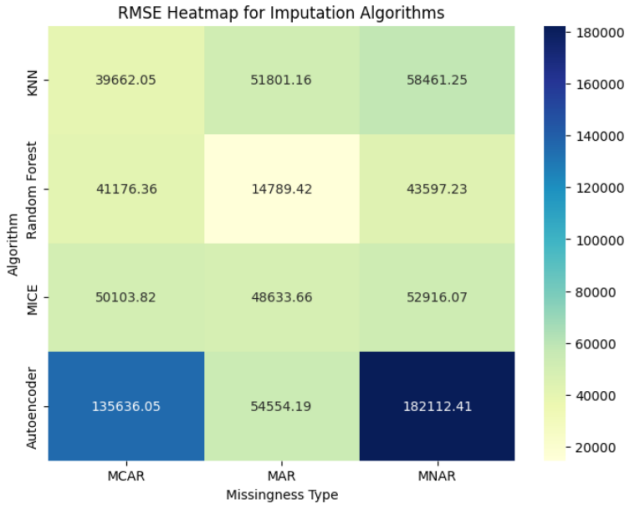

fig. 3

6. DISCUSSION

RMSE heatmaps combined with time efficiency data show that each imputation method has its own strengths depending on the type of intrusion.

MCAR: Random Forest and KNN are the most accurate and efficient, with Random Forest providing a good balance between accuracy and computation time.

MAR: Random Forest is the best in terms of accuracy, but Autoencoder is also efficient but time consuming. MICE and KNN remain practical when computing needs are low.

MNAR: Autoencoder is the most accurate, but is computationally expensive, making it impractical for large data sets. Random Forest and MICE provide reasonable accuracy with better performance for MNAR.

This analysis highlights the tradeoffs between imputation accuracy and time efficiency. Random forests and Autoencoders are generally preferable for precision in MAR and MNAR, while KNN and the MICE offer effective alternatives for MCAR scenarios or when temporary restrictions are necessary.

Table 1

	MCAR	MAR	MNAR
KNN	6.1764	0.4157	0.8507
Random Forest	21.121	4.4411	3.9798
MICE	0.2951	0.2828	0.2040
Autoencoders	20.2011	23.0134	21.3819

time taken

Table 2

	MCAR	MAR	MNAR
KNN	39662	51801	58461
Random Forest	41176	14789	43597
MICE	50103	48633	52916
Autoencoders	135636	54554	182112

root mean squared error

7. CONCLUSION

This study evaluated the effectiveness of the four attribution methods: Autoencoders, KNN method, random forest, MICE, various types of missing data mechanisms, MCAR, MAR, and data with MNAR. Each imputation technique is evaluated based on two major measurements: average two major measurements (RMSE), which measure the accuracy and time of calculation indicating efficiency. The results indicate that the choice of imputation method should be determined by the nature of the missing data and the specific trade-off between accuracy and computational resources.

In summary, the findings reveal that:

- Random Forest is a reliable and accurate choice for both MCAR and MAR data, and it performs reasonably well for MNAR.
- KNN is suitable for MCAR scenarios where efficiency is a priority over high accuracy.
- MICE is versatile across all types of missing data, balancing accuracy and efficiency, making it a good general-purpose imputation method.
- Autoencoders are highly accurate for MAR and MNAR but are computationally intensive, making them best suited for cases where accuracy is paramount and computational resources are available.

8. FUTURE SCOPE

Hybrid Imputation Methods: Merging various imputation techniques may improve performance by utilizing the advantages of each algorithm. For instance, applying Autoencoders for the initial imputation and then utilizing MICE for refinement could enhance both precision and efficiency.

Hyperparameter Optimization: Adjusting the hyperparameters of each imputation approach, especially for machine learning methods such as Random Forest and Autoencoders, could enhance accuracy and decrease computation time further.

Application to Varied Datasets: Evaluating these methods on datasets from different fields (such as healthcare, finance, social sciences) with unique structures and relationships may enhance our comprehension of their practical use in real-world scenarios.

Investigation of Deep Learning Approaches: Sophisticated deep learning frameworks such as GANs (Generative Adversarial Networks) and Transformers, designed for data imputation, may be examined for their ability to manage intricate missingness patterns.

9. REFERENCES

[1] Sun, Y., Li, J., Xu, Y., Zhang, T., Wang, X. (2015). Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications, Volume 227, 0957-4174. https://doi.org/10.1016/j.eswa.2023.120201

[2] Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

[3] Huang, J., Xu, L., Qian, K., Wang, J., Yamanishi, K. (2021) Multi-label learning with missing and completely unobserved labels. Data Min Knowl Disc **35**, 1061–1086. https://doi.org/10.1007/s10618-021-00743-x

[4] Lin, W., Tsai, C., Zhong, J. (2022) Deep learning for missing value imputation of continuous data and the effect of data discretization, Knowledge-Based Systems 239. https://doi.org/10.1016/j.knosys.2021.108079