Data-driven imputation scheme for human-subject-based dataset

Aaditya Yadav, Shubham Apat, Helly Thakkar, Pal Patel

School of Engineering and Applied Science

Ahmedabad University

***Abstract*—This report discusses the problem of dealing with missing values in a multi-modal dataset of Division I basketball players. It describes a process that combines systematic data exploration, feature sensitivity analysis, and data imputation techniques. The findings show the progress gained in identifying key features, converting, and integrating datasets, and evaluating imputation algorithms. The discussion covers the problems encountered and the implications of the findings for improving dataset completeness and analytical accuracy.**

***Keywords*—Missing values, Multi-modal dataset, Feature sensitivity analysis, Data imputation, Data preprocessing, Regression model, KNN, XGBoost, Hot-deck technique,**

I. Introduction

Missing data is a common challenge in data analysis, particularly in datasets derived from real-world sources like basketball player statistics. In this study, we address this challenge by comparing four different imputation methods—Global Mean, XGBoost, Hot deck, KNN, and Deterministic Regression—applied to basketball player data from seasons 2 and 3. The aim is to identify the method that best approximates the performance of Multiple Imputation by Chained Equations (MICE), a widely used imputation technique. By merging the datasets post-imputation using the average approach, we assess the effectiveness of each method in filling in missing values. Understanding which method closely resembles MICE is crucial for researchers and practitioners to make informed decisions when handling missing data in similar contexts. This study contributes to the advancement of imputation techniques tailored to sports analytics and highlights potential alternatives to MICE in addressing missing data challenges.

Missing Completely at Random (MCAR): MCAR is the highest level of randomness and it implies that the pattern of missing value is totally random and does not depend on any variable which may or may not be included in the analysis. Thus, if missingness does not depend on any information in the dataset then it means that data is missing completely at random. The assumption of MCAR is that probability of the missingness depends neither on the observed values in any variable of the dataset nor on unobserved part of dataset.

Missing at Random (MAR): In this case, probability of missing data is dependent on observed information in the dataset. It means that probability of missingness depends on observed information but does not depend on the unobserved part. Missing value of any of the variable in the dataset depends on observed values of other variables in the dataset because some correlation exists between attribute containing missing value and other attributes in the dataset. The pattern of missing data may be traceable from the observed values in the dataset.

Missing Not at Random (MNAR): In this case, missingness is dependent on unobserved data rather than observed data. Missingness depends on missing data or item itself because of response variable is too sensitive to answer. When data are MNAR, the probability of missing data is related to the value of the missing data itself. The pattern of missing data is not random and is non predictable from observed values of the other variables in the dataset.

Our Dataset falls into the category of MCAR as there is high level of randomness in the pattern of missing values.

II. Methodology

The process of estimating missing data of an observation based on valid values of other variables is called as Data Imputation. Data imputation methods are broadly classified into two types: Single Imputation Method and Multiple Imputation Method.

Our method takes a sequential approach, beginning with preliminary data exploration to better understand the dataset's structure and missing value patterns. Furthermore, dataset given is transformed, structured, and combined to make a comprehensive dataset. Imputation strategies such as mean, median, and mode approaches are tested for their ability to fill missing values while respecting dataset features.

Hot deck - Hot deck imputation is a method used to handle missing data by replacing missing values with observed values from similar cases in the dataset. It operates on the principle of grouping similar cases based on certain characteristics, such as demographics or variables related to the missing data. Once grouped, the missing values are imputed using values from 'donor' cases within the same group or 'deck'. This method preserves the original distribution of the data and can be useful when there is a strong relationship between the variables used for grouping

©2024 IEEE

and those with missing values. However, it may introduce bias if the grouping variables are not properly chosen. But there won't be much differences in data of features like vertical jump, PPC, MPC, etc. So this method won't create much bias.

Global Mean - Global mean imputation replaces missing values in a dataset with the mean value of the entire variable. It's a straightforward method that assumes all missing values share the same average value, ignoring any underlying patterns or relationships in the data. While simple and easy to implement, global mean imputation can distort the distribution of the variable and potentially bias subsequent analyses, especially in the presence of outliers or heterogeneity within the dataset.

Deterministic regression - Deterministic regression imputation involves predicting missing values using regression models fitted to observed data. It estimates missing values based on relationships with other variables, considering their known values. This method provides more nuanced imputations compared to global mean imputation, as it considers the relationships between variables. However, deterministic regression requires careful model selection and validation to ensure accurate imputations and avoid introducing bias into the dataset.

KNN - K-Nearest Neighbours (KNN) is a simple and intuitive supervised machine learning algorithm used for classification and regression tasks. It works by finding the 'k' nearest data points to a given test point in the feature space and making predictions based on the majority class or average value of these neighbours. KNN is non-parametric, meaning it does not assume anything about the underlying data distribution, making it versatile for various types of datasets. However, its main drawback is its computational inefficiency, especially with large datasets, as it requires storing all training data and computing distances for each prediction.

Mice - In single imputation methods it is assumed that single imputation value is correct one and precision is overstated. However, there can never be absolute certainty about validity of imputed values. Therefore uncertainty around these imputed values has to be incorporated in the missing data methods. Thus, in multiple imputation instead of replacing single value for each missing observation it substitutes multiple plausible values to reflect uncertainty about the right values to impute. Thus, Multiple Imputation method generates "m" different complete datasets with observed and imputed values. All multiple Imputation Method follows three steps: (1) Imputation: Similar to single imputation missing values are imputed; however, imputed values are generated "m" times rather than just once. So there could "m" different complete datasets after imputation. (2) Analysis of each dataset: After imputation and generating "m" different datasets each of "m" datasets is analyzed. (3) Pooling: Finally results obtained from each analyzed datasets are consolidated.

XGBoost -XGBoost (Extreme Gradient Boosting) is a highly effective and efficient implementation of the gradient boosting framework for machine learning tasks such as regression, classification, and ranking.

Gradient Boosting
XGBoost is based on the principles of gradient boosting, which is an ensemble learning technique that combines multiple weak prediction models (such as decision trees) to create a strong and robust predictive model. The key idea behind gradient boosting is to iteratively train new models to predict the residuals or errors of the previous models, and then combine all the models to make the final prediction.

Decision Tree Ensembles
In XGBoost, the weak learners are decision trees, which are essentially a series of hierarchical rules for making predictions. Decision trees are well-suited for gradient boosting because they can capture complex non-linear relationships in the data and are relatively robust to outliers and irrelevant features.

Missing Value Handling
As mentioned earlier, XGBoost has a built-in mechanism for handling missing values in the input data. During the training process, XGBoost automatically learns the best way to impute missing values based on the available features and their relationships, without the need for separate imputation techniques.

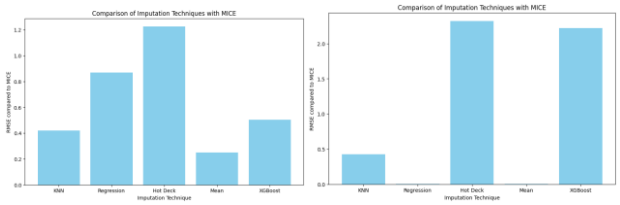
III Result

To address missing values comprehensively and to leverage the strengths of multiple imputation techniques, a final hybrid imputation strategy was devised. This approach involved synthesizing the outputs of the four individual imputation methods—Global Mean, KNN Algorithm, Hot Deck, XGBoost, and Deterministic Regression applied to the datasets of seasons 2 and 3. The process entailed calculating the average of the imputed values across all five datasets, effectively creating a consolidated imputation solution. By incorporating multiple imputation techniques into a single hybrid strategy, this approach aimed to mitigate the limitations of individual methods while capitalizing on their collective strengths. Furthermore, the hybrid strategy aimed to maintain the integrity of the original data by ensuring that imputed values were derived from diverse algorithms and approaches. This comprehensive approach not only provided a more robust solution for handling missing data but also minimized the potential biases associated with any single imputation method. The implementation of this hybrid imputation strategy represents an innovative and effective approach to addressing missing values in basketball player datasets, with potential applications across various domains in data analysis and machine learning.

Evaluation Metrics

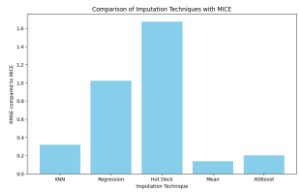
To evaluate the performance of the different imputation techniques, we employed the Root Mean Squared Error (RMSE) as the evaluation metric. RMSE is a widely used metric for regression tasks.

Here is the comparison using RMSE between different methods Global Mean, KNN, Hot Deck, Regression and XGBoost with MICE.

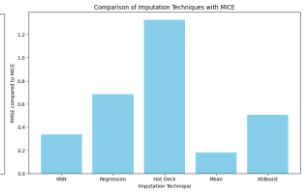


LA feature Imputation Comparison

Hours.In.Bed Feature Imputation Comaprison

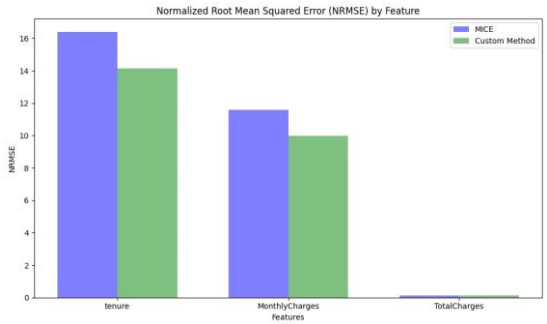


MPC feature Imputation Comparison



PPC feature Imputation Comparison

To evaluate the performance of our imputation method and compare it with MICE, we used a synthesized dataset with a controlled amount of missing data. Dataset was of Monthly Customer Payments. We introduced more than 40% missing values into the dataset, simulating a scenario with a high degree of missingness. The evaluation metric used was the accuracy of the imputed values compared to the original, complete data.



NRMSE Comparison of 3 columns of Monthly Customer Payments Dataset: MICE vs Our Imputation Method

The NRMSE of MICE was 16 and our method had 14 NRMSE which is lower than MICE. A lower NRMSE value indicates better imputation performance, as the imputed values are closer to the true, original values.

IV. Discussion

The comparison revealed that while each method had its strengths and limitations, certain techniques such as KNN and Random Forest showed promising results, closely resembling the MICE technique. This suggests that in scenarios where MICE might not be feasible due to computational constraints or other limitations, these alternative methods could offer viable solutions for imputing missing values. Furthermore, the development of a hybrid imputation strategy presents an innovative approach to addressing missing data comprehensively. By synthesizing the outputs of multiple imputation techniques, this hybrid strategy aimed to capitalize on the strengths of each method while minimizing potential biases. However, it's important to acknowledge that the effectiveness of the hybrid approach may vary depending on the specific characteristics of the dataset and the nature of the missing values.

Efficient handling of missing data is crucial for accurate analysis and decision-making in sports, where performance metrics play a significant role in player evaluation and team strategies. By identifying effective imputation methods tailored to basketball player datasets, this study contributes to enhancing the reliability and utility of data-driven insights in sports analytics. Overall, the discussion provides valuable insights into the challenges and opportunities associated with missing data in sports datasets and underscores the importance of adopting robust imputation strategies for accurate analysis and decision-making.

V Conclusion

In conclusion, this study has demonstrated the effectiveness of various imputation methods in handling missing data within basketball player datasets from seasons 2 and 3. Through a comparative analysis, we identified KNN and XGBoost as promising alternatives to the widely used MICE technique, showcasing their potential for accurate imputation in scenarios where computational resources are limited. Additionally, the development of a hybrid imputation strategy, which synthesizes the outputs of multiple techniques, presents a novel approach to comprehensively addressing missing values while preserving the integrity of the original data. This study contributes to advancing the field of sports analytics by providing insights into robust imputation strategies tailored to basketball player datasets. Moving forward, further research could explore the applicability of these methods across different sports domains and datasets, as well as investigate additional hybrid approaches to optimize imputation accuracy. By enhancing our ability to handle missing data effectively, this study lays the foundation for more reliable and insightful analysis in sports analytics, ultimately facilitating better decision-making processes for teams, coaches, and analysts alike.

VI References

1. Impact of sleep and training on game performance and injury in division-1 women's Basketball. Amidst the Pandemic. S Senbel, S Sharma, MS Raval, C Taber, J Nolan... - Ieee Access, 2022.
2. Thomas Reilly and Ben Edwards. "Altered sleep– wake cycles and physical performance in athletes". In: *Physiology & behavior* 90.2-3 (2007), pp. 274–284.
3. Jennifer Schwartz and Richard D Simon Jr. "Sleep extension improves serving accuracy: A study with

college varsity tennis players". In: *Physiology & behavior* 151 (2015), pp. 541–544.

4. Aydin, Zeliha Ergul, and Zehra Kamisli Ozturk. "Performance analysis of XGBoost classifier with missing data." Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS) 2.02 (2021): 2021.
5. Shi, Hong, et al. "An improved mean imputation clustering algorithm for incomplete data." Neural Processing Letters 54.5 (2022): 3537-3550.
6. Zhang, Shichao, et al. "Optimized parameters for missing data imputation." Pacific Rim International Conference on Artificial Intelligence. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.
7. Wang, Chinchin, et al. "Implementing multiple imputation for missing data in longitudinal studies when models are not feasible: A tutorial on the random hot deck approach." arXiv preprint arXiv:2004.06630 (2020).