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**.**

II. Methodology

The method takes a sequential approach, beginning with preliminary data exploration to better understand the dataset's structure and missing value patterns. Feature sensitivity analysis is used to prioritize features that are critical for imputation. Furthermore, data from various sources 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 and those with missing values. However, it may introduce bias if the grouping variables are not properly chosen.

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 - Multiple Imputation by Chained Equations (MICE) is a method used to handle missing data in datasets. It operates by imputing missing values through an iterative process where each variable with missing data is imputed using the observed values from other variables. MICE leverage the relationships between variables to impute missing values multiple times, hence the term 'chained equations'. This approach is advantageous as it preserves the structure and variability of the data while imputing missing values, providing more accurate and reliable results compared to simple imputation techniques.

XGBoost - XGBoost, short for Extreme Gradient Boosting, is a powerful and efficient implementation of the gradient boosting framework, which is widely used for supervised learning tasks such as classification, regression, and ranking. XGBoost builds a series of decision trees sequentially, where each subsequent tree corrects the errors made by the previous ones. It incorporates regularization techniques to prevent overfitting and utilizes a gradient-based optimization algorithm to minimize a predefined loss function. XGBoost is known for its speed, scalability, and state-of-the-art performance, often winning machine learning competitions and being extensively adopted in industry settings for its robustness and accuracy.

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 five 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.

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).