IJIRIS: International Journal of Innovative Research in Information Security
Volume 09, Issue 03, June 2023
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives

Addressing the challenges of handling missing data in Data science applications

Prof Shobha V
Asst. Professor, Department of CSE
Sri Sairam College of Engineering, Anekal, Bengaluru, India
shobav.cse@sairamce.edu.in
Mahima Sahu, Mahesh Kumar Chaudhary, Manu Priya K
UG Scholars, Department of CSE
Sri Sairam College of Engineering, Anekal, Bengaluru, India
sce20cs104@sairamtap.edu.in, sce20cs027@sairamtap.edu.in
sce20cs091@sairamtap.edu.in,

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.09/Issue03/JNIS10084
Research Article | Open Access | Double-blind Peer-reviewed |Article ID: IJIRIS/RS/Vol.09/Issue03/JNIS10084
Received: 02, June 2023 | Accepted: 20, June 2023 | Published: 23, June 2023
Volume 2023 | Article ID JNIS10084 http://www.ijiris.com/volumes/Vol09/iss-03/05.JNIS10084.pdf
Article Citation: Shoba,Mahima,Mahesh,Manu(2023)Addressing the Challenges of Handling missing data in Data science applications. International Journal of Innovative Research in Information Security (IJIRIS), 10, 76-78.
doi: https://doi.org/10.26562/ijiris.2023.v0903.05
BibTex key: Shobha2023address

Copyright: ©2023 This is an open access article distributed under the terms of the Creative Commons Attribution License; which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Abstract: Missing data is a common issue in data science applications that can significantly impact the accuracy and reliability of data analysis and modeling. This paper explores the challenges associated with missing data and presents various approaches and techniques to effectively handle missing data in data science applications. We discuss both traditional statistical methods and advanced machine learning techniques that can be employed to address missing data. Additionally, we highlight the importance of understanding the underlying mechanisms causing missingness and the implications it has on data analysis. Through this paper, we aim to provide a comprehensive overview of the existing methods and methods for dealing with missing data, offering insights into their practical applicability and future research directions in this field.

Keywords: Missing data, Data science, Data analysis,Data modeling, Data quality, Data preprocessing, Machine learning techniques, Mecahnism for missing data.

I.INTRODUCTION

In the realm of data science applications, the accessibility to complete and reliable data plays a vital role in accurate analysis and robust modeling. However, that they are presence of missing Data is a pervasive challenge that can significantly hinder the effectiveness and validity of data-driven endeavors. Missing data alludes to the absence of values or observations in a database. The impact of data analysis and data mining of missing data and modeling is multifaceted. It can lead to biased estimates, reduced statistical power, and diminished generalizability of findings. Additionally, errors in data can distort relationships between variables, affect model performance, and impede the discovery of meaningful patterns or insights. This is how the paper is organised: Paragraph2 provides a comprehensive overview of the types and mechanisms of missing data. In Section 3, we discuss traditional statistical methods commonly employed to handle missing data. Section 4 delves into the most recent Machine learning techniques. In conclusion, addressing the challenges posed by by lacking data is of utmost importance in data science applications. By employing suitable techniques and methodologies, researchers and practitioners can overcome the limitations caused by missingness, leading to more accurate analyses, reliable models, and enhanced decision-making processes. Through this paper, we aim to provide a comprehensive resource that facilitates a deeper understanding of the challenges associated with missing data and equips data scientists with the necessary knowledge to effectively address this issue.

II - TYPES AND MECHANISM OF DATA MISSING

The types of missing data can provide insights into the underlying mechanisms causing the missingness. The several types of Missing data: Missing fully at Random (MCAR), Missing at Random(scar), and absent but not Random(MNAR).

You can imagine these as traffic lights MCAR is green (easy), MAR is amber(intermediate) and MNAR is red(problematic))[1].

1. Missing completely all at once

Missing entirely and arbitrarily (MCAR) is a situation where how likely it is that the information is neither a specified value nor missing are connected, which is intended to be acquired, or the collection of seen responses [2].

2. Missing at random

MAR, or missing at random, is a realistic assumption. It is unrelated to the particular variables that are absent but should be obtained [2].

3. Missing not at random

If the data's characters don't match the requirements for MCAR OR MAR, then they fit this description not missing at random [2].

III - TRADITIONAL METHODS

Now that we have learnt what is missing information and its types we need to learn about the methods used for dealing with the incomplete data. There are two types of methods used, that is, Traditional statistical method and Modern methods based on machine language and artificial intelligence. The traditional statistical methods commonly employed to handle missing data, including deletion by list and pair, imputation of the mean, and regression imputation, and multiple imputation.

1. Listwise and Pairwise deletion
   Listwise deletion, sometimes referred to as a full case study or casewise deletion, is a method used to handle missing data in statistical analysis. When a dataset contains missing values, listwise deletion involves removing any cases or observations that have missing data on any of the variables included in the analysis[2]. Pairwise deletion refers to the practice of excluding observations with missing data on specific variables from calculations involving those variables while including them in calculations involving other variables.[2].
2. Mean imputation
   Mean Impugnation is a typical method used in dealing with lacking values in a data set, a dataset are replaced,having an average value of the available information regarding that variable. It is a straightforward technique that aims to preserve the overall mean of the variable while making up the omitted values[3].
3. Regression Imputation
   Regression impugnancy is a method used when it comes to missing data analysis to estimate missing values derived from relationships observed in the available data. It involves using regression models to predict using different variables to fill in the blanks that are available in the dataset[4].
4. Multiple Imputation
   Multiple imputation aims to address the issue of ignoring the incomplete cases by imputing, or filling in, the missing values based on the observed values and the relationships within the dataset. The process involves creating multiple plausible imputed datasets, where missing values are replaced with estimated values based on statistical models.[5].

To overcome the shortcomings of traditional statistical methods we have developed newer and better techniques using Machine language and artificial intelligence. Those methods are discussed in the next Section.

IV MODERN METHODS

The previous Section discussed traditional methods used in handling missing data. Some of the modern methods of machine learning, like Expectation-Maximization (EM) algorithm, K-nearest neighbors (KNN) imputation, matrix factorization, deep learning approaches, and Bayesian methods.

1. K-nearest neighbors (KNN)
   KNN imputation is a technique that fills in missing values by estimating them based on the values of their K-nearest neighbors[6].
2. Matrix factorization
   The basic idea behind matrix factorization is to approximate the incomplete matrix as the product of two lower-rank matrices. This decomposition allows us to capture the latent structure or underlying recurring themes in the data, which then be utilised to determine the error value[7].

IJIRIS © 2014-23, AM Publications -All Rights Reserved

3. Expectation-Maximization algorithm
   It is particularly useful when lacking data mechanism classified as "missing completely at random" (MCAR) or "missing at random" (MAR). The EM algorithm aims to locate the maximum likelihood estimates (MLE) of the settings when missing data is present[8].
4. Bayesian methods
   It is a statistical approach based on the principles of Bayesian inference, it uses probability theory to quantify and update beliefs about parameters or in light of new evidence. It provides a powerful framework for addressing missing data in data science applications. These methods incorporate prior knowledge and uncertainty in a principled manner, allowing for more robust and flexible imputation of missing values.

V - CONCLUSION

Handling missing data is a critical challenge in data science applications, as missing values can significantly impact the validity and reliability of the analysis. In this paper, we have explored missing data and its types and also discussed handling techniques for it. We have seen that missing data can occur in different patterns, such as MCAR, MAR, or MNAR, each requiring specific treatment strategies. Traditional statistical methods, multiple imputation techniques, and machine learning approaches have been extensively studied and applied to impute missing values. Additionally, Bayesian methods offer a powerful framework that incorporates prior knowledge and uncertainty to address missing data. The choice of imputation technique should align with the specific characteristics of the dataset, the missing data pattern, and the goals of the analysis. Addressing the challenges of handling missing data is an ongoing area of research, and future work should focus on developing innovative techniques that can handle complex missing data scenarios. In conclusion, the challenges of handling missing data in data science applications require careful consideration and appropriate methodologies. By employing the right imputation techniques and accounting for the uncertainty associated with missing values, researchers can mitigate the biases and limitations caused by missing data and improve the accuracy and reliability of their analyses. Addressing the challenges of missing data is essential for ensuring the validity of results and facilitating robust decision-making in data-driven applications.

REFERENCES

[1]. Missing Data: Introducing the Missingness Mechanism- Tessa Wilkie, Lancaster University.
[2]. The prevention and handling of the missing data- Hyun Kang.
[3]. Handbook of Latent Variable and Related Models, 2007- Mortaza Jamshidian, Matthew Mata.
[4]. Applied missing data analysis with spss and r studio- Heymans and Eekhout.
[5]. Multiple Imputation: A Flexible Tool for Handling Missing Data- Peng Li, Elizabeth A. Stuart and David B. Allison.
[6]. KNNImputer: A robust way to impute missing values (using Scikit-Learn)- Kaushik Roy Choudhary.
[7]. Matrix Factorization Explained- Great Learning Team
[8]. Complete Guide to Expectation-Maximization Algorithm- Chirag Goya.