Approach to handle missing values in different Datasets

Dr.Manisha Bharati
Professor, Department of technology
Savitribai Phule Pune University
Pune,India

Rutuja Awalkar
Department of technology
Savitribai Phule Pune University
Pune,India

Niranjan Kadam
Department of technology
Savitribai Phule Pune University
Pune,India

Abstract

In a well-designed and controlled observations, missing values occurs in almost all data. There are many types of studies for which missing values are an issue. Missing values can reduce the statistical power of a study and can make influence on valid conclusions. This study reviews the types and techniques for handling missing data. The methods by which missing data types are illustrated, and the methods for handling the missing values are discussed. The paper concludes with recommendations for the handling of missing data.

Introduction

Missing data is an everyday problem that a data professional need to deal with. What is missing data? Missing data are defined as not available values, and that would be meaningful if observed. Missing data can be anything from missing sequence, incomplete feature, data entry error, files missing, information incomplete, etc. Most datasets in the real world contain missing data. Before you can use data with missing data fields, you need to transform those fields to be used for analysis and modelling. Like many other aspects of data science, this too may actually be more art than science. Understanding the data and the domain from which it comes is very important. Having missing values in a data is not necessarily a setback. Still, it is an opportunity to perform the right feature engineering to guide the model to interrupt the missing information right way.

A. Types of Missing Data

The crucial part of the pre-processing of data is to deal with the missing values, which may be prevalent due to one of the following reasons:

1) Missingness completely at random: It can be said that data is missing completely at random if the probability for all missing values is equal. When we say data are missing completely at random, we mean that the missingness is nothing to do with the subject of observation.

2) Missingness at random: The probability that values are missing depends on available information. In this case, process can be modeled as logistic regression with outcome 1 for available and 0 for missing data. Missing data can be treated as NA.There is a relationship between the propensity of a value to be missing and its values. In other words, data are missing not at random when the missing values of a variable are related to the values of that variable itself, even after controlling for other variables.

3) Missingness that depends on unobserved variables: Data is no longer considered to be randomly missing if it is dependent on variable that have not been recorded. When we say data are missing at random, we mean that the missingness is to do with the object of observation but can be predicted from other information about the person. It is not specifically related to the missing information

B. Methods for handling missing values

1.*Replacing missing values with zero*

In these cases, missing values are replaced by 0. Sometimes the missing values represent nothing then it is safe to replace it with 0.

2.*Replacing missing values with mean*

In this technique goal is to replace missing data with statistical estimates of the missing values. Mean can be used as imputation value. In a mean substitution, the mean value of a variable is used in place of the missing data value for that same variable. This has the benefit of not changing the sample mean for that variable. The theoretical background of the mean substitution is that the mean is a reasonable estimate for a randomly selected observation from a normal distribution. However, with missing values that are not strictly, especially in the presence of great inequality in the number of missing values for the different variables the mean substitution method may lead to inconsistent bias. Distortion of original variance and distortion of co-variance with remaining variables within the dataset are two major drawbacks of this method.

3.*Replacing missing values with median*

In the cases with skewed data distribution and outliers present, median imputation is used over mean imputation.

4.*Replacing missing values with mode*

The third measure of central tendency that can be used is mode. Mode imputation is appropriate method for missing nominal data and fairly competent method for numerical data.

5.*Replacing missing values with forward fill*

If data is time-series data, one of the most widely used imputation methods is the last observation carried forward (LOCF). Whenever a value is missing, it is replaced with the last observed value. This method is easy to understand and communicate.

6.*Replacing missing values with backward fill*

A similar approach like LOCF works oppositely by taking the first observation after the missing value and **c**arrying it backward**.**

7.*Linear regression*

Variable prediction with missing values is identified using a correlation matrix. In regression equation the best predictors are opted and considered as independent variables. The variable having missing data is treated as the dependent variable. for generating regression equation, the cases with complete data for the predictor variables are used; then the missing values for incomplete cases is predicted by equation. In an iterative process, values for the missing variable are inserted and then all cases are used to predict the dependent variable.

Results

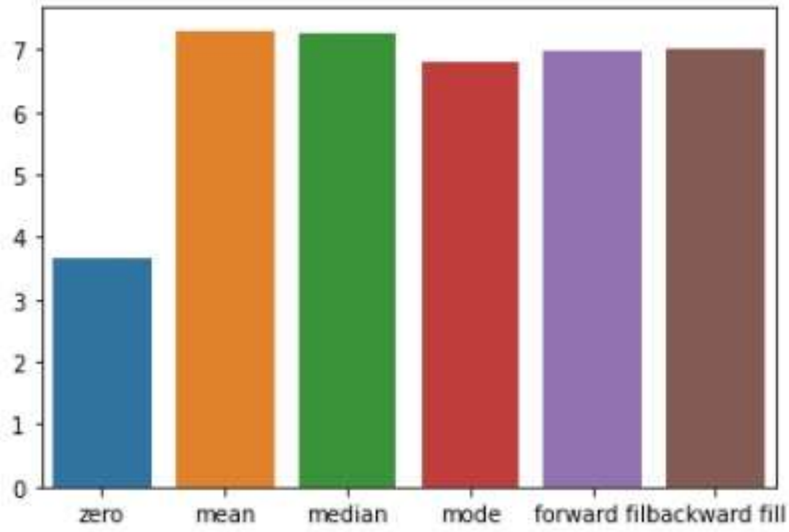

Fig 1. House-Price Dataset results

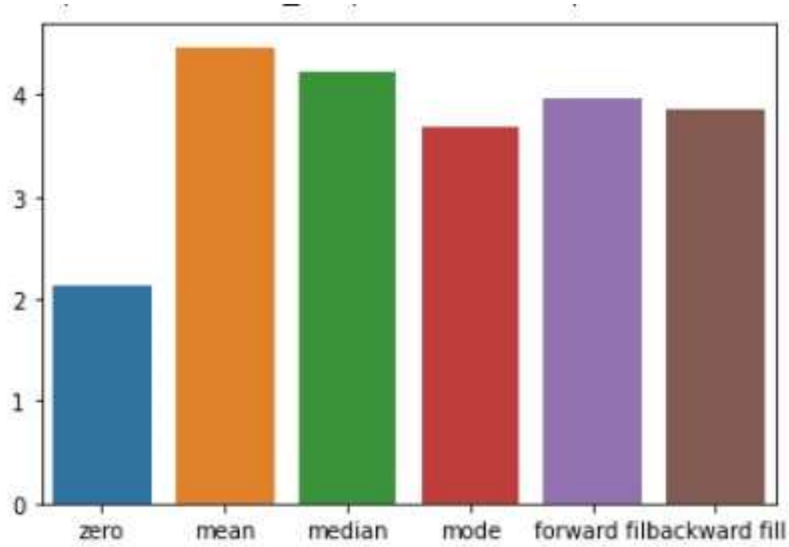

Fig 2. Melbourne dataset results

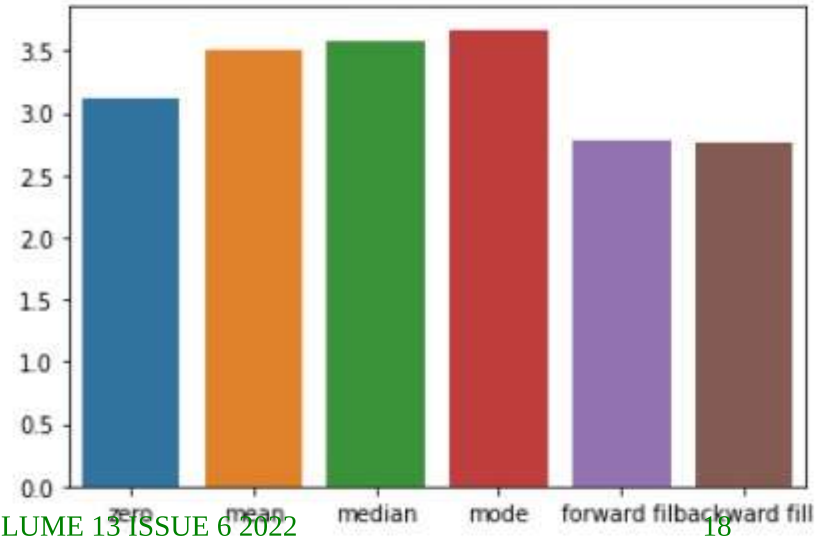

Fig 3. Titanic dataset results

Fig 1, Fig 2 and Fig 3 shows the different methods applied on house price dataset, Melbourne dataset and Titanic dataset respectively. The results after applying different methods in these datasets conclusion can be made that mode and mean are most efficient methods to handle the missing values.

Reference

1. E. Acuna and C. Rodriguez. The treatment of missing values and its effect on classifier accuracy. In Classification, Clustering, and Data Mining Applications, pages 639–647. Springer, 2004.

2. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol.* 2009; 60:549–576.

3. G. Batista and M. Monard. K-nearest neighbor as imputation method. Technical report, Experimental Results. Tech. Report 186, ICMC-USP, 2002.

4. K. Baumgartner, S. Ferrari, and G. Palermo. Constructing Bayesian networks for criminal profiling from limited data. Knowledge-Based Systems, 21(7):563–572, 2008.

5. 17. Acock AC. Working with missing values. *J Marriage Fam.* 2005; 67:1012–1028.

6. Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. Annual Review of Public Health, 25, 99–117.

7. Acock, A. C. (1989). Measurement error in secondary data analysis. In K. Namboodiri & R. Corwin (Eds.), Research in sociology of education and socialization (Vol. 8, pp. 201–230). Greenwich, CT: Jai Press.