International Journal of Research in Engineering, Science and Management
Volume-2, Issue-4, April-2019
www.ijresm.com | ISSN (Online): 2581-5792

Missing Data Analysis using Artificial Intelligence

Shreya Banerjee[1], Shikha Mishra[2], Harshada Pai[3]
[1,2,3]*Student, Department of Electronics and Telecommunications Engineering, Thakur College of Engineering and Technology, Mumbai, India*

***Abstract*: One of the biggest problems faced in data observation or data recording process is the frequent occurrence of missing values. The needs for data completeness of the observation data for the uses of advanced analysis becomes important to be solved. Conventional method such as mean and mode imputation, deletion and other methods are not good enough to be handling missing values as those methods can cause bias to the data. Some of the procedures or algorithms are best solved using Estimation or imputation to minimize the use of conventional methods for finding the missing data. So that at last, the data will be complete and ready to use for another step of analysis or data mining. It is the process of estimating missing values in time series data for unilabiate data involves methods such as analysis and modelling. The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.[5][8]**

***Keywords*: Missing data, database, parameters, imputations, models, Rstudio.**

1. Introduction

Future 500 is a database which includes a yearly data of 500 of the largest US companies ranked by total revenues for their respective fiscal years. The list is compiled using the most recent figures for revenue and includes both public and private companies with their respective revenue data. The survey includes parameters like unique identification (ID) of that company, name of the company, its respective industry, its inception , the no. of employees that are incorporated, the state in which the company is situated, its city ,the annual revenue of the company , the expenses , the amount of profit and the growth of the company in that particular year. Sometimes, some of the data of each of these parameters get missed. Due an error in the functionality during the data collection, also at times due to human error all the data does not get updated on a yearly basis. Due to lack of any one of parameter, estimation of the revenue and expenses of the particular company gets troublesome to obtain. This leads to a problem in determining the profit of that company. And if profit is not known finding the growth of the company through all these years becomes a major issue. Each parameter in the database plays a very crucial role. So missingness of any of the data does cause a very serious problem. Hence to solve the problem due to missingness of a particular data, we analyze the data using missing data analysis [1].

2. Motivation

As we know, with even the slightest change in the data for large corporations can create large impact in country's annual revenue leading to economic disruption. With the involvement of the analysis for missing data aid to bridge the issues related to data imputation. This study establishes an approach for the further study of prediction to improvise these important parameters related to specific companies This study involves the methodology relating to the precise attainment of data on a marginal scale with the purpose for future extension [2], [5].

3. Objectives

Description: We can use the missing data analysis methods to obtain illustrative or statistical measures of a time series. For example, here to measure various parameters relating to specific company, you can plot a graph. We can measure the profit and the growth in the past few years and thus find out a solution to maintain or improve the accountability it in the near future.

- *Explanation*—we can use the observed variation of a time series to explain the variation of a related time series, which can help you understand the nature of the relationship between the two parameters. For example, we can use this to compare the results obtained by various models.
- *Prediction*—With the observed data over the years, using factual imputation the city can be predicted and by using mean and mode imputation we can predict the revenues, expenses and then by using formula imputation profit can be evaluated with respect to expenses and expenses and according to profit the growth can be predicted.
- *Control*—as we have already collected the observed values and also predicted the future values and know about the seriousness of it, we can find out the parameters that influence the growth of company. Once we have obtained it, now we can do the required analysis, to what extent is each of the parameters

contributing in the growth and profit of company and further find ways to maintain the profit for that company [6].

4. Methodology

Different parameters are treated differently. This database consists of parameters like the (ID) of that company , name of the company, its respective industry , its inception , the no. of employees that are incorporated, the state in which the company is situated, its city ,the annual revenue of the company , the expenses , the amount of profit and the growth of the company in that year. So different type of imputations is used to find the missing data in any parameters of a particular database. There are different imputations such as

- Factual Imputation,
- Mean and Mode Imputation
- Formula Imputation

A. Factual analysis method

Correcting our dataset with factual analysis. Restore data with 100 % certainty. Example: If we know the city we can find the state as well and be 100% sure about the restored values. fin[is.na(fin$State), & fin$City == "New York", "State"] = NY – The above line first looks for any NA in the state column using is.na on State, and (&) then looks for City which are equal to 'New York' and then replaces the NA 'State' with our assigned notation of state New York i.e. NY. This method does the replacement of all the New York city and updates the NA in the state of equivalent row as NY [4], [5].

B. Median imputation method

We can use median imputation method for example says number of employees: if we don't know the number of employees for certain company than we may take median and of all the company's employees in dataset and assign the median value in the missing values. Further scrutiny can also be done with using more filters say, company in same sector, state, city, etc. looking for reality. Mostly use Median as it does bypass outliers. Median (fin[fin$Industry == "Retail", "Employees"], na.rm = TRUE) – This gives us the median of employees in Industry for Retail sector. Furthermore, it is saved as med_empl_retail. na.rm is used to bypass NAs in the employee's column. fin[is.na(fin$Employees) & fin$Industry == "Retail", "Employees"] = med_empl_retail– This replaces the employee value which was as NA to the median value stored in med_empl_retail [4].

C. Formula imputation

Some NAs can be factually replaced with real values using relevant formulas. For example, Profit NA can be replaced using formula Revenues – Expenses which is factually correct, and we can be 100% sure about the same. Fin [is.na (fin$Profit), "Profit"] = fin[is.na(fin$Profit), "Revenues"] – fin [is.na(fin$Profit), "Expenses"] This firstly looks or is any NA (is.na) in here 'Profit' column and replaces the NA with difference of Revenue and Expenses. Here, say if any other of 3 Expenses, Revenue or Profit is missing can be replaced in same fashion [4], [1]. Data might be not collected, lost while transferring, etc. are the possible reasons for the loss of data i.e. missing data.

- Predict missing data with 100% accuracy. Example: State and City, Revenue Expenses and Profit
- Leave record as it is Example: Field not important for analysis, algorithm (advance) we use takes care for the missing data
- Remove record entirely Example: to make algorithm working removed entire field
- Replace with a mean or median (mostly used) Example: Missing revenue data in field can be replaced using mean or a median, median is mostly preferred over mean as it is less affected by the outliers.
- Fill in by exploring correlations and similarities Example: Create Regression or models which makes model predict what could be the possible data to fill in
- Introduce dummy variable for 'Missingness' (…...?) Example: Introduce a variable which raises 'yes' flag for missing and 'no' flag for not missing and explore correlation with to the outcome you are looking for [4], [8].

5. Algorithm

- *STEP 1:* Set the working directory of your datasheet in Rstudio.
- *STEP 2:* Read the datasheet by first converting your datasheet by adding an extension ".csv" after the file name.
- *STEP 3:* In order to know which are the different parameters, use tail and specify how much rows and columns you want. In order to know where your data is ending use the command tail with specified rows and columns no.
- *STEP 4:* The missing values in different parameters are treated as NA
- *STEP 5:* changing from factor to a non-factor
- *STEP 6:* Pattern matching and replacement sub() and sub() These functions help us remove unwanted things from the field and substitute it with the new one as per desired. sub() function does it only for the first one in the field while gsub() does it for the entire field.
- *STEP7:* Dealing with Missing Data different imputations are performed
- *STEP 8:* Locate missing values

6. Software used

R is an open-source programming language and environment alternative to S. It is mainly used for statistical computing and graphics. R mainly has a command-line interface but has many

graphical front-end interfaces, like the popular RStudio. It is also accessible from various scripting languages like Python, Perl, Ruby etc. is mainly used by different branches of data science professionals, mostly data miners. Knowing how to work with R is beneficial for many sub-fields of Data Science and required in some, e.g.- Data Analysis, Machine Learning, Data Visualisation, Statistical Modelling etc. There is a fair amount of demand for professionals of Data Science in India and the predictions are that the demand is going to increase as the companies realize the value of data analysis. There are many job opportunities in startups and any growing company. Data analyst is something for which almost every growing company is ready to hire.The Data Science field is very wide and has a lot of potential for growth [5].

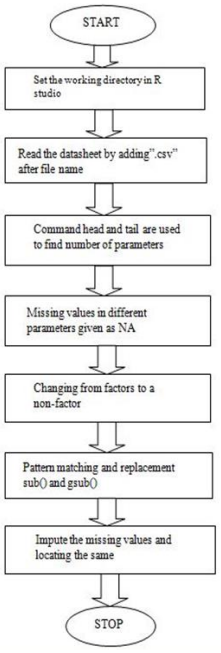

Fig. 1. Flow chart

7. Result & discussion

	ID	Name	Industry	Inception	Employees	State	City
1	1	Over-Hex	Software	2006	25	TN	Franklin
2	2	Uninattax	IT Services	2009	36	PA	Newtown Square
3	3	Greenfax	Retail	2012	NA	SC	Greenville
4	4	Blacklane	IT Services	2011	66	CA	Orange
5	5	Yearflex	Software	2013	45	WI	Madison
6	6	Indigoplanet	IT Services	2013	60	NJ	Manalapan
7	7	Treslam	Financial Services	2009	116	MO	Clayton
8	8	Rednindox	Construction	2013	73	NY	Woodside
9	9	Lantone	IT Services	2009	55	CA	San Ramon
10	10	Stripfind	Financial Services	2010	25	FL	Boca Raton
11	11	Canecorporation	Health	2012	6	<NA>	New York

Above are the missing values denoted as<NA>

	ID	Name	Industry	Inception	Employees	State	City
1	1	Over-Hex	Software	2006	25	TN	Franklin
2	2	Uninattax	IT Services	2009	36	PA	Newtown Square
3	3	Greenfax	Retail	2012	28	SC	Greenville
4	4	Blacklane	IT Services	2011	66	CA	Orange
5	5	Yearflex	Software	2013	45	WI	Madison
6	6	Indigoplanet	IT Services	2013	60	NJ	Manalapan
7	7	Treslam	Financial Services	2009	116	MO	Clayton
8	8	Rednindox	Construction	2013	73	NY	Woodside
9	9	Lantone	IT Services	2009	55	CA	San Ramon
10	10	Stripfind	Financial Services	2010	25	FL	Boca Raton
11	11	Canecorporation	Health	2012	6	NY	New York

The Missing values <NA> are found using imputation method.

8. Conclusion

Missing values can be treated in much way dependence on the issue caused by the missing values. The method of missing values can affect the outcome of the analysis. There are various ways to find the missing data like mean imputation, mode imputation, factual analysis and formula imputation with the help of R software. The parameters in the dataset are independently linked to each other. Missingness of any one of the parameter can cause issues in the calculation of profits being evaluated for the company. Hence the method proposed above can help to eradicate such issues that occurs while data observation and recording processes.

Acknowledgement

We sincerely thank our guide Dr. Sangeeta Mishra for her guidance and constant support and also for the stick to our backs. We also thank the project coordinators for arranging the necessary facilities to carry out project work. We thank the HOD, Dr. Vinit Kumar Dongre, Dean Academic, and Dr. R. R. Sedamkar, Vice Principal, Dr. Deven Shah, The Principal, and Dr. B. K. Mishra and the college management for their support.

References

[1] Percival, Donald B.; Walden, Andrew T. (1993). Spectral Analysis for Physical Applications. Cambridge University Press.
[2] Bower man, B.L. and R.T. O'Connell, Time Series Forecasting, Duxbury Press, Boston MA.
[3] Box, G.E.P. and G.M. Jenkins, Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, CA, 1976.
[4] R M Bhardwaj etc.; Status of different companies, New York, April, 30, 2015.
[5] Jui-Sheng Chou, Chia-Chun Ho, Ha-Son Hoang, 2017. Estimation using r-studio.
[6] Misaghi, F., Delgosha, F., Razzaghmanesh, M. & Myers, B. Introducing a water quality index for assessing water for irrigation purposes: a case study of the ghezel ozan river. Science of the Total Environment 589, 107–116, 2017.
[7] WHO, guidelines for drinking water quality world health organization, Geneva, Switzerland 2008.