iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VI **Month of publication:** June 2018

DOI: http://doi.org/10.22214/ijraset.2018.6101

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

Data Exploration and Preparation

K. Rushikesh[1] P. Harika Reddy[2]
[1]*Student, Bml Munjal University, Gurugram, India*

Abstract: Missing data is the most important problem in large data sets. Even we use simple techniques to solve missing data, the focus of this paper is to explain each and every steps in data cleaning and visualization. Many techniques were developed in to solve missing data and visualize the data. This paper concentrates on identifying the different types of variables, Data types , category, Techniques used in finding them, strategies of treating missing values. This analysis we have carried out using different examples and explained.

I. INTRODUCTION

We need to prepare a dataset before we actually implement the machine learning algorithm on the dataset.Data is the raw information it is the representation of both humans and machines observation of the world. Everything can be represented as the data for example science, literature, Art everything in the world can be transformed as the data by transforming into ones and zeroes by the computer. Once we actually decide the data set used to solve our problem now we start our data exploration and preparation.

Data exploration is an main approach to data analysis where the person uses exploration to understand what actually is present in the data set and what are the variables present and the we would actually come to know the characteristics of the data set.Data exploration actually helps us in selecting the right tool for pre processing or analysis.Successful visualisation requires data to be converted into a visual format.by data visualisation the data characteristics and relation between data items will be clearer and can be reported on or analysed faster and more efficiently.

Data set may contain discrepancies in the names may contain out liners or errors. There may be lack of attributes. To overcome all these problems we need to data exploration and we need to prepare the data for analysis. The Quality of data exploration give us the most quality input which results in getting the most quality Output.

II. LITERATURE REVIEW

[10] Little and rubin summarize the imputation methods. Also introduces mean imputation methods find out the missing values. The drawbacks of mean imputation are size is overestimated, variance is underestimated. For median and standard deviation also replacing all missing records with single value will deflate the variance and artificially inflate the significance if any statistical tests based on it.[11] Roderick Little and Donald Rubin summarize statistical analysis with the missing data and points to estimate methodologies for handling missing data.And also consider main area of statistics:Study of Variance,Multivariate Analysis and Survey sampling[12].charu C.Aggarwal gives abstract of elementary algorithms for Outlier analysis completely and covering advanced data types Understanding when the algorithms work productively.[13]

III. IDENTIFY TYPES OF VARIABLES,DATA TYPE,CATEGORY

Target variable	Predictor Variable	Category-Continuous	Category-Categorical
Bike Buyer	Marital Status	Yearly Income	Marital Status
	Gender	Children	Gender
	Yearly Income	Cars	Education
	Children	Commute Distance	Occupation
	Education	Age	Home Owner
	Occupation		Region
	Home Owner		

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VI, June 2018- Available at www.ijraset.com

	Cars		
	Commute Distance		
	Region		
	Age		

IV. UNIVARIATE ANALYSIS

For continuous variables understand Central tendency and spread of variable using Mean, Median, Mode, Min, Max, Standard deviation, Variance. Perform data visualization on independent variables using Histogram and Box Plot.
For Categorical variables use frequency table to understand the category of each variable using count and count%. Use Bar chart for visualization.

V. BIVARIATE ANALYSIS

Finds out relation between two variables. Checks association and dis association between variables at a predefined Significance level.
Can be performed for combination of
→Categorical and Categorical(Use two way table , stacked column chart, Chi-Square Test)
→Categorical and Continuous(Z-Test/T-Test/Anova)
→Continuous and continuous(Use correlation and scatter plot)

VI. STRATEGIES USED TO TREAT MISSING VALUES

A. Deletion

it's of 2 types: List Wise Deletion and Pairwise Deletion.

B. List Wise Deletion

In list wise deletion, we tend to delete observations wherever any of the variable is missing. Simplicity is one in all the most important advantage of this technique, however this technique reduces the ability of model as a result of it reduces the sample size.

C. Pairwise Deletion

In pairwise deletion, we tend to perform analysis with all cases within which the variables of interest. Advantage of this technique is, it keeps as several cases accessible for analysis. one in all the disadvantage of this technique, it uses totally different sample size for various variables.

VI. DATA EXPLORATION MISSING VALUES AND DELETION STRATEGIES

Deletion strategies are used once the character of missing knowledge is “Missing completely at random” else non random missing values will bias the model output.

A. Mean/ Mode/ Median Imputation:

Imputation could be a technique to fill within the missing values with calculable ones. the target is to use illustrious relationships that may be known within the valid values of the information set to help in estimating the missing values. Mean / Mode / Median imputation is one in all the foremost oftentimes used strategies.
It consists of substitution the missing knowledge for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all illustrious values of that variable. It may be of 2 types Generalized Imputation:during this case, we tend to calculate the mean or median for all non missing worths of that variable then replace missing value with mean or median.
Similar case Imputation:
In this we calculate the average of the attribute of the non missing values individually and then replace the missing values based on the attributes.

B. Prediction Model

Prediction model is one in all the subtle technique for handling missing knowledge. Here, we tend to produce a prognostic model to estimate values that may substitute the missing knowledge. during this case, we tend to divide our knowledge set into 2 sets: One set with no missing values for the variable and another one with missing values. 1st knowledge set become coaching knowledge set of the model whereas second knowledge set with missing values is take a look at knowledge set and variable with missing values is treated as target variable. Next, we tend to produce a model to predict target variable supported different attributes of the coaching knowledge set and populate missing values of take a look at knowledge set.We can use regression, ANOVA, logistic regression and numerous modeling technique to perform this.

C. There are Pair of Drawbacks for This Approach

1) The model estimated values are sometimes a lot of well-behaved than truth value
2) If there aren't any relationships with attributes within the knowledge set and also the attribute with missing values, then the model won't be precise for estimating missing values.
3) KNN Imputation: In this technique of imputation, the missing values of associate degree attribute are imputed using the given variety of attributes that are most kind of like the attribute whose values are missing. The similarity of two attributes is determined using a distance function.advantage & disadvantages of the KNN function are

D. Advantages

1) k-nearest neighbour will predict each qualitative & quantitative attributes
2) Creation of prognostic model for every attribute with missing knowledge isn't needed
3) Attributes with multiple missing values may be simply treated
4) Correlation structure of the information is taken into thought

E. Disadvantage

1) KNN algorithm is much time consuming in analysing large database.Because the algorithm searches the whole dataset for the similar instances or groups.
2) Choice of k-value is extremely essential. Higher worth of k would come with attributes that are considerably totally different from what we want whereas lower worth of k implies missing out of serious attributes.

After managing missing values, future task is to manage outliers. Often, we tend to tend to neglect outliers whereas building models. this is often a discouraging follow. Outliers tend to form your knowledge skew and reduces accuracy. Let's learn a lot of concerning outlier treatment. The most commonly used method to detect outliers is visualization.We use various visualization methods like Boxplot,Scatter Plot,Histogram

VII. TECHNIQUES OF OUTLIER DETECTION AND TREATMENT

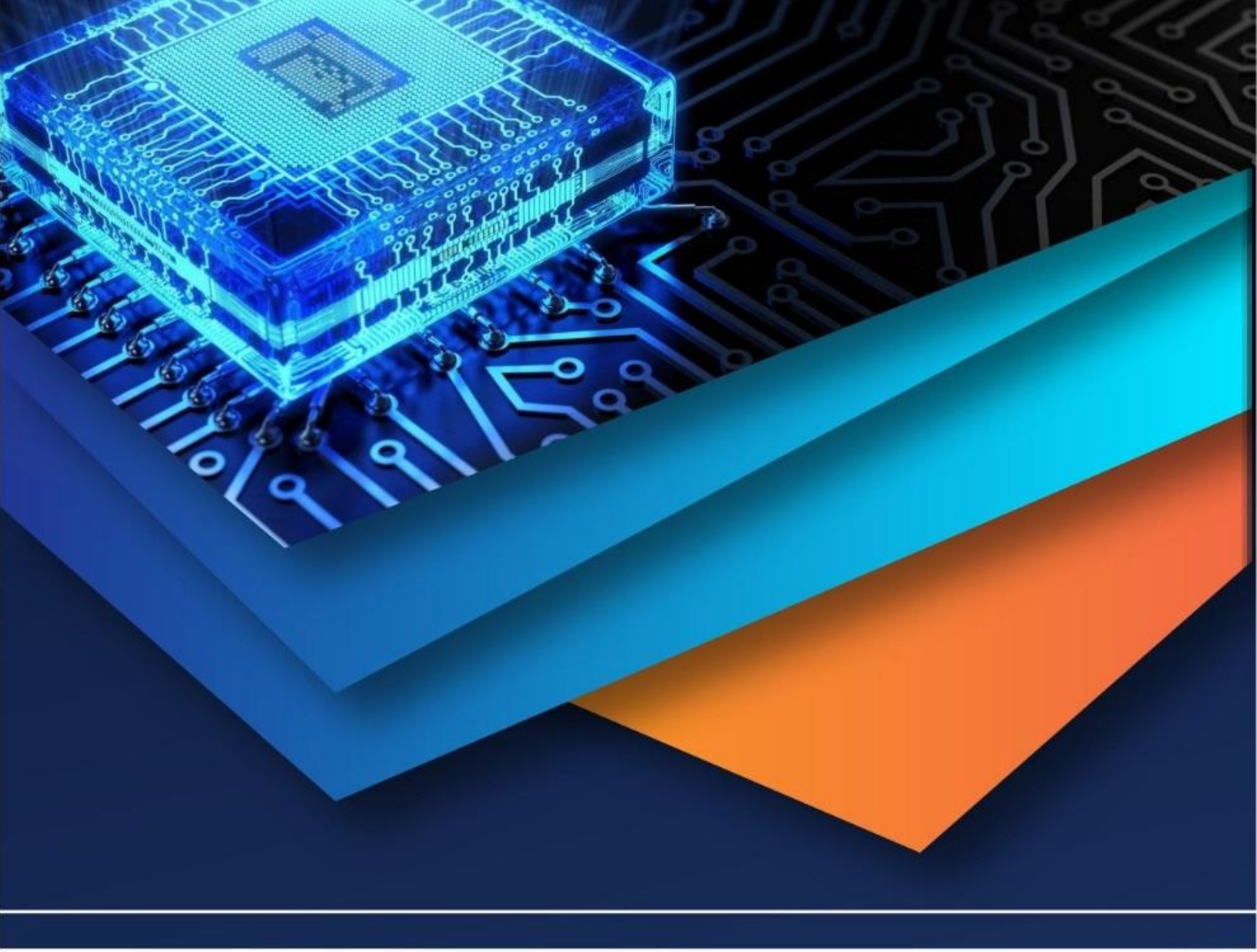
General rules for deleting Outliers
⇨Any value which is beyond the range of -1.5*IQR TO 1.5*IQR should be deleted.
⇨Any value which is out of range of 5th and 95th percentile should be considered as outlier.
⇨Data points,three or more standard deviation away from mean are considered as outliers.
If an outlier is artificial then remove it by mean/median/mode imputing
To remove outliers
→Delete Observations
→Transform
→Binning
→Imputing Values
→Other statistical Methods

VIII. CONCLUSIONS

This paper gives the complete view about the multiple imputation of missing values in the large data set analysing them and visualising them. The paper is mainly focused on data exploration and preparation and techniques in solving and preparing the dataset. Strategies used in treating missing values. Outlier techniques and detection and treatment, prediction models and KNN imputation.

REFERENCES

[1] Ray, Sunil, and Business Analytics. "A Complete Tutorial Which Teaches Data Exploration in Detail." Analytics Vidhya, 2 May 2017, www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/
[2] Data Exploration and Preparation." Data Exploration and Preparation - Data Science Studio 2.0.0 Documentation, doc.dataiku.com/dss/2.0/preparation/index.html.
[3] Top 38 Data Preparation Tools and Platforms - Compare Reviews, Features, Pricing in 2018 -PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices."Predictive Analytics Today, Bigtexts.com, 22 May 2017, www.predictiveanalyticstoday.com/data-preparation-tools-and-platforms/
[4] How to Handle Missing Data with Python." Machine Learning Mastery, 10 Mar. 2018, machinelearningmastery.com/handle-missing-data-python/
[5] How To Handle Missing Values In Machine Learning Data With Weka." Machine Learning Mastery, 22 June 2016, machinelearningmastery.com/how-to-handle-missing-values-in-machine-learning-data-with-weka/.
[6] Replacing Missing Value by Class Conditional Mean." Data Science Stack Exchange, datascience.stackexchange.com/questions/23603/replacing-missing-value-by-class-conditional-mean
[7] Maladkar, Kishan. "5 Ways To Handle Missing Values In Machine Learning Datasets."Analytics India Magazine, 12 Feb. 2018, analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/
[8] How to Identify Outliers in Your Data." Machine Learning Mastery, 7 June 2016, machinelearningmastery.com/how-to-identify-outliers-in-your-data/.
[9] Casas, Pablo. "Outliers Treatment." Outliers Treatment · Data Science Live Book, livebook.datascienceheroes.com/data_preparation/outliers_treatment.html.
[10] 3 Methods to Deal with Outliers." Neural Designer | Advanced Analytics Software, www.neuraldesigner.com/blog/3_methods_to_deal_with_outliers
[11] B.Rabbit, et al. "What Is the Point of Univariate and Bi-Variate Analysis?" Data Science, Analytics and Big Data Discussions, 5 Aug. 2016, discuss.analyticsvidhya.com/t/what-is-the-point-of-univariate-and-bi-variate-analysis/11004.
[12] What Is Multivariate Analysis? • r/MachineLearning." Reddit, www.reddit.com/r/MachineLearning/comments/2mrgax/what_is_multivariate_analysis/.
[13] Exploratory Data Analysis." Wikipedia, Wikimedia Foundation, 9 June 2018, en.wikipedia.org/wiki/Exploratory_data_analysis
[14] Mitchell, Tom M. "Machine Learning (Mc-Graw Hill - Tom Mitchell, 1997) by - DBLab by Tom M. Mitchell - PDF Drive." Free PDF Drive to Download Ebooks., www.pdfdrive.net/machine-learning-mc-graw-hill-tom-mitchell-1997-by-dblab-e15568857.html
[15] VanderPlas, Jake. "Python Data Science Handbook." Introducing Scikit-Learn | Python Data Science Handbook, jakevdp.github.io/PythonDataScienceHandbook/
[16] Replacing Missing Value by Class Conditional Mean." Data Science Stack Exchange datascience.stackexchange.com/questions/23603/replacing-missing-value-by-class-conditional-mean
[17] What Is Multivariate Analysis? • r/MachineLearning." Reddit, www.reddit.com/r/MachineLearning/comments/2mrgax/what_is_multivariate_analysis/
[18] R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY