# Chapter 14
# Missing Data Handling by Mean Imputation Method and Statistical Analysis of Classification Algorithm

**K. Maheswari, P. Packia Amutha Priya, S. Ramkumar, and M. Arun**

## 14.1 Introduction

Nowadays data mining tools are used for analysis of customer behavior and their relationship to increase the profit by several companies such as retail marketing, insurance, banking, telecommunications and product sales of consumer, finance and health care. To understand the business aspects, data mining helps the organization to provide betterment of customer satisfaction and serve in order to increase the growth of the organization in the future.

To increase the market space among other competitors, the retailers can know all the information related to the customer based on why, what and they buy the products and who the customer are. By applying the benefits of data mining techniques to various sectors, the huge quantities of data related to customer behavior, product supplier, list of products, and their sales of product were analyzed.

In any marketing business, data mining plays an important role not only in paying more attention on customer but also in maintaining the existing customers without leaving the competitors. In the retail marketing business, information gained by data mining techniques can be useful in various ways:

- Profit of the business can be increased by reducing the cost of the product.
- Stock price can be increased in order to maintain the future plan to raise the profit.

If a marketing business fails to retain the existing customers, then the company will not be able to retain its position in the market, their shares will go down and the profit of the company goes down slowly and it finally disappears. This chapter

K. Maheswari (✉) · P. Packia Amutha Priya · S. Ramkumar · M. Arun
School of Computing, Kalasalingam Academy of Research and Education, Virudhunagar, India

A. Haldorai et al. (eds.), *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-19562-5_14

is organized as follows: Section 14.2 deals with review of literature, Sect. 14.3 describes the methodology, Sect. 14.4 presents the results of this work, and Sect. 14.5 concludes the research work.

## 14.2 Review of Literature

Pratama et al. [1] discussed about the reasonable option used for guessing out the missing values used by other researchers in various research work and suggested that mean, mode, median, deletion, and other imputation methods are methods used to handle missing values in various time series dataset. The imputation method used in the research work will produce a solution to minimize the preference outcome of the predictable technique.

Moore and Carpenter [2] suggested decision tree method to analyze the behavioral and demographic issues on private label attire from the top five private label attire merchant from the USA. The author pointed out observable drivers is more frequent among customers of vendors that are distinguished based on brand or service provided to the customer.

Maheswari and Packia Amutha Priya [3] analyzed the purchase behavior of the customer using SVM algorithms and the experimental result is carried out using the inventory dataset and sales dataset for analyzing the customer purchase behavior. Maheswari and Packia Amutha Priya [4] experimented the research work using text documents. After preprocessing, the missing values in the documents were found out. The frequency of words present in the documents is analyzed and visually represented. The twitter dataset [5] was used to carry out classification using SVM and KSVM. This work focuses on categorizing the sentiments or emotions from various age groups of people. The author [6] experiments the twitter dataset using Knn to identify the human behavior by improving the accuracy result in the sentiment classification analysis.

Tinabo [7] described four possible data mining methods to the problem of customer retention in the retail sector and suggested decision tree to be the most effective technique. The decision is made based on the features of the retail datasets such as size of records. Islam and Habib [8] proposed a prediction model for analyzing the business region in retail commercial banking. Business customer records of both rural and urban fields from Bangladesh dataset is taken for the experiment by applying decision tree method in weka tool to analyze its performance.

Patel et al. [9] suggested building a decision tree classification model to categorize the training dataset based on the rules. Result of the proposed classification model strength was calculated based on their performance. Li and Zhang [10] surveyed the solution for handling empty value property, selecting more than one value property and condition based selection property problems. Simplified and weighted Entropy in decision tree algorithm performs better results when compared to ID3 algorithm during the experimental process.

Senapati et al. [11] mentioned a principal component analysis model to clear up the missing values present in the microarray dataset. The result shows that the accuracy produced by the proposed model was good when compared with other imputation methods. Linear Discriminant Analysis (LDA) was used to validate the proposed model.

Houard et al. [12] recommended the new sampling methodology to handle missing values during preprocessing process. The main aim was to produce the samples of good trait and the extracted information should be highly secure and dependable. Houcka et al. [13] inspects about the way to categorize the missing values depending on the machine learning and statistical techniques. In this work, the classification of the missing values by imputation, model-based, and ignoring value methods was performed. Each of the methods has integrated with their own merits and demerits.

Song and Lu [14] suggested visualizing the training dataset results in tree structure in order to characterize the SAS and SPSS programs by applying various algorithms namely QUEST, CHAID, CART, and C4. 5. Validation dataset is used to determine the correct tree size and attain the excellent model validation dataset which was used for their analysis.

Agarwal et al. [15] aimed to categorize the society college dataset of the student using support vector machines. Various classification methods are compared for their research study and find that SVM produces more accuracy and less root mean square error. Decision tree method may be used to find the course selection of the students during the program. Kishor Kumar Reddy et al. [16] attempted to summarize the proposed approaches, tools, etc. for decision tree learning with emphasis on optimization of constructed trees and handling large datasets. Further, we also discussed and summarized various non-decision tree approaches like Neural Networks, Support Vector Machines, and Naive Bayes.

## 14.3 Methodology

It is a diagrammatic flow structure where each node represents an attribute; each branch is a result of test condition. The leaf or terminal nodes represent a class label. The top node in a decision tree is designated as root node [17]. The decision tree is useful for constructing decision tree classifiers without any basic knowledge. It can be able to handle higher dimensional data in a database. Good accuracy can be given by decision tree classification technique [18].

### *14.3.1 Methods for Handling Missing Data*

Missing data or values become one of the major problems that occur frequently during the data collection process. Missing data reduces the sample representation and there was a alter presumption in population. It is important to know the reason why data values are missing in order to correct the remaining data. Missing data can be handled in different ways:

- Missing at Random (MAR)
- Missing completely at Random (MCAR)
- Missing not a Random (MNAR)

#### 14.3.1.1 Missing at Random

It happens when the absence of data was not random or unplanned whereas the absence of data can be entirely computed for some variable where the information are complete.

#### 14.3.1.2 Missing Completely at Random (MCAR)

Missing Completely at Random (MCAR) means any particular data values being missed are separate for unobserved variables and observed variables of important and occur fully at random choice.

#### 14.3.1.3 Missing Not a Random (MNAR)

Non-ignorable non-response is also known as Missing not a Random (MNAR) is data point or values that is neither MCAR nor MAR. The missing value depends on the two probable reasons such as missing data is dependent on few other variable values or hypothetical value.

If an observation has one or more missing values, then all data value can be removed. This is known as list wise deletion. To a fewer number of observation, if the missing value is finite, then it is preferred to ignore these cases from the analysis process. Imputation is the process of replacing the missing values by some other predictable values in the entire dataset. It is the important step in the machine learning and data mining process whenever the definite values are missed in the Dataset. There are various ways to handle missing data in dataset:

- Removing the entire row in the dataset.
- Replacing the missing value with mean, median, and mode in the dataset.
- Assigning a unique categorical value.
- Missing values can be predicted.

Dataset is divided into two sets. They are:

- Without missing values as a one set of data for the training set.
- Dataset with missing values for the testing set has been used the work.

Using decision tree, logistic regression and ANOVA method for prediction process in retail marketing has been used in our proposed work and it is implemented by R programming.

## 14.4 Methodology Experimental Results

Missing data is an important issue in datasets which takes main role in statistical processing of machine learning algorithms. The data missed is not by the mistake of data collector. There are so many reasons to miss the data in the dataset. During data collection, the respondent may not respond for some questions. Sometimes the data will not be available, not applicable, and not possible to provide. The lack of non-responded answers is treated as missing values. The proper handling of the missing values, empty values, NA values, not relevant values, and improper values draws inaccurate conclusion about the data set. One of the most important techniques for handling missing data is imputation method.

It can be seen that the variables shown in the above chart have missing values from 30% to 40%. The margin plot is shown in Fig. 14.1.

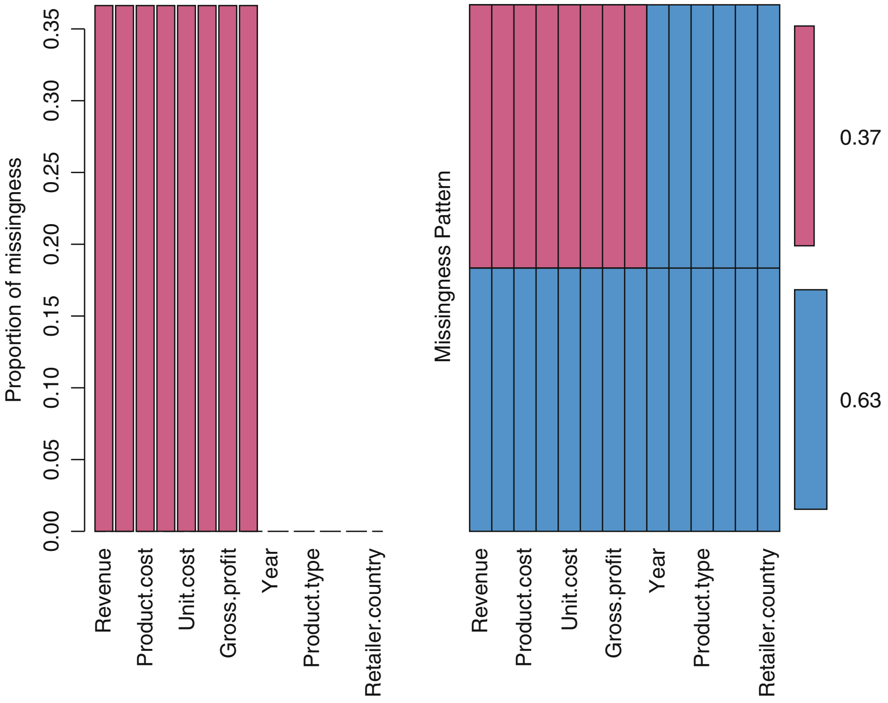


**Fig. 14.1** Dataset with missing values

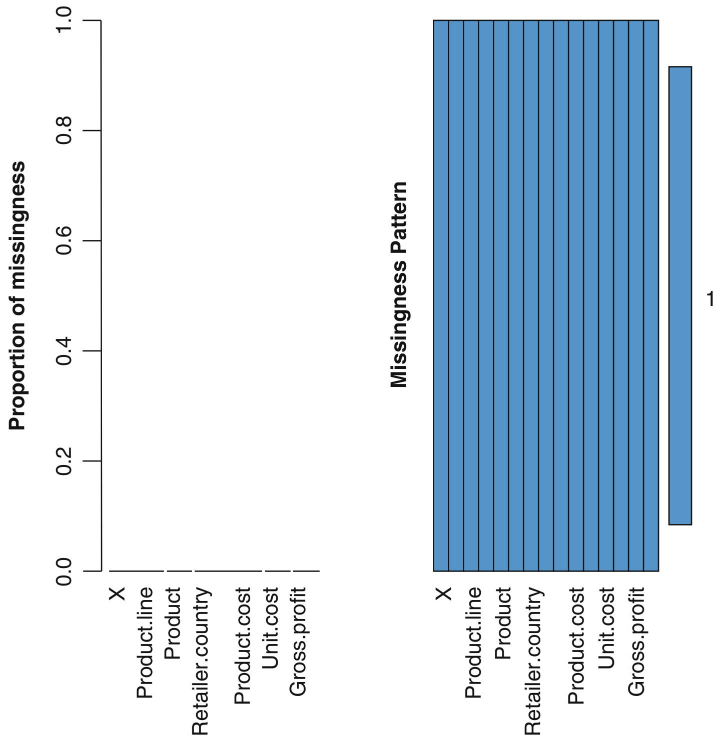


**Fig. 14.2** Missing pattern

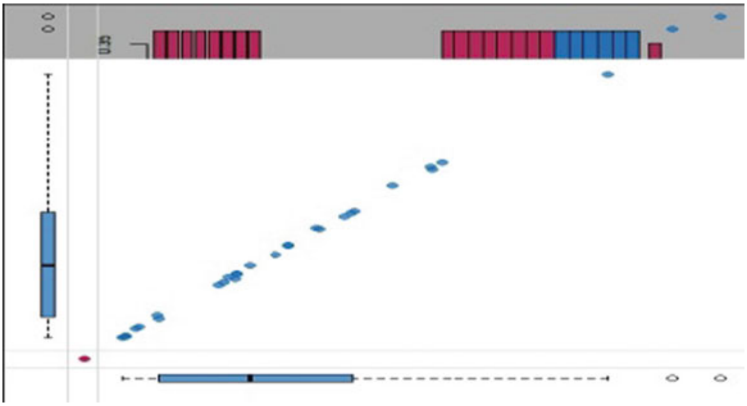


**Fig. 14.3** Margin plot of dataset

The missing pattern of dataset is presented in Fig. 14.2. The margin plot of data set is shown in Fig. 14.3. The plotting of two attributes, revenue and product cost, is shown in Fig. 14.4.

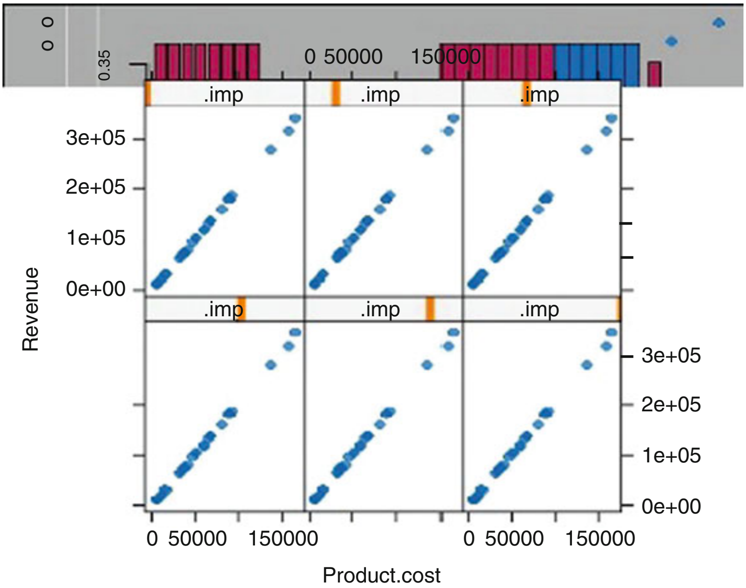


**Fig. 14.4** Plot with two attributes

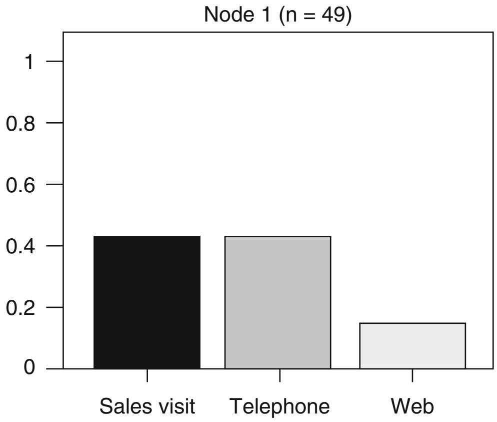
The margin plot indicates the plotting of two attributes, revenue and product cost, at a time. The blue color indicates the experiential data. The red one represents the mean imputed data. The red plot indicates missing distribution of one feature. The blue box is the distribution of attributes which is present. The red box plot and blue box plot are equal, the missing values in the data set is MCAR (Missing Continuously At Random). The above chart represents that the values are not missing Continuously At Random. Hence mean imputation is performed for missing values. The next thing is, measuring whether the imputed value is good or bad. The xyplot() and densityplot() functions are used to plot, compare, and verify our imputations. The sales order method type is plotted in Fig. 14.5.

The number of samples used in this work is 49. There are three types of order method type used in this sample.

- Sales visit to a particular site
- Through telephone
- Through internet or web

From these observations, it is found that the sales order method type 1 and 2 were used equally which is 40% whereas type 3 (Through Internet) was used by people which are less than 20%. The retail marketing data set is downloaded from the internet and 50 samples were taken into consideration. The data present in the dataset is both numeric and nonnumeric. The classification algorithm rpart is used for classification with order method type and unit sale price of mean imputed data and is shown in Fig. 14.6. The retailer country is a feature in this data set and is

**Fig. 14.5** Sales order method type plot



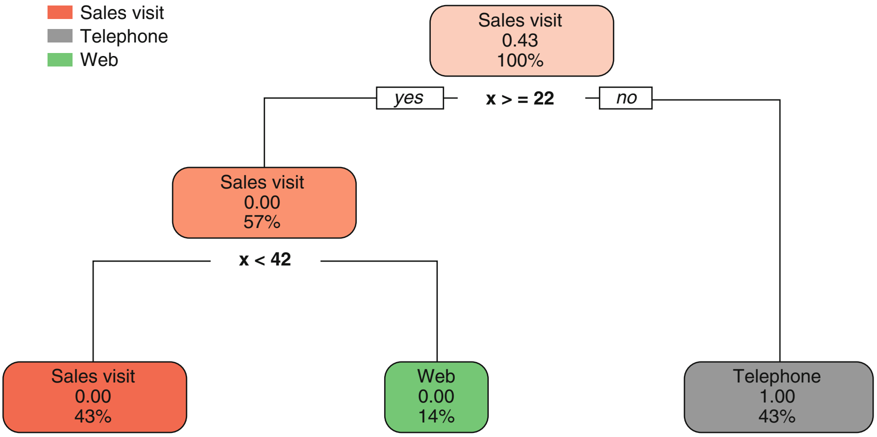


**Fig. 14.6** Decision tree for order method type

used for classification and is shown in Fig. 14.7. The plot of unit sale price among countries is shown in Fig. 14.8.

GLMs are most commonly used to model binary data or countable data to predict a class accurately. The output of the function lies between 0 and 1. The glm model is built with two attributes, namely order method type and retailer country of mean imputed data with binomial family link logit.

The general linear model of glm is

$$\hat{Y} = \beta 0 + \beta 1X \tag{14.1}$$

where

$\hat{Y}$ is predicted variable,
$\beta 0$ constant value,

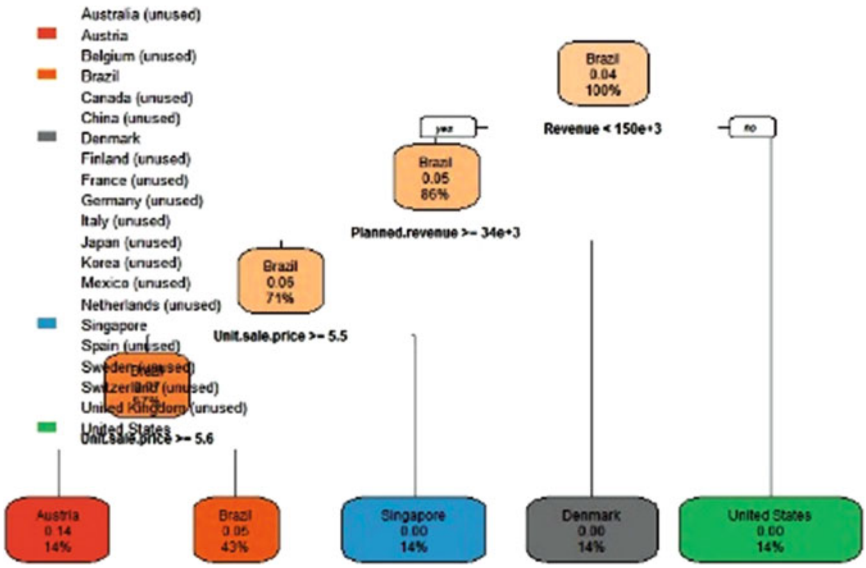
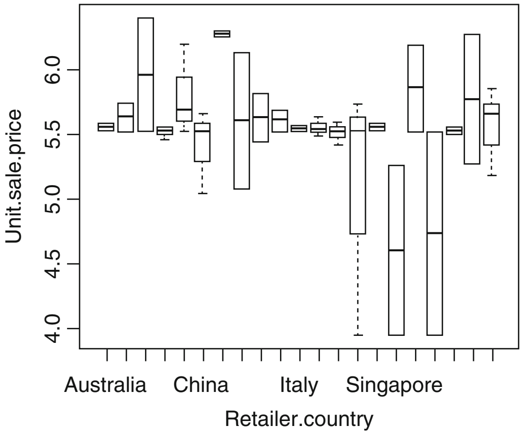


**Fig. 14.7** Retailer country tree

**Fig. 14.8** Plot of unit sale price among countries



$\beta 1$ coefficient weight, and
$X$ is variable

The glm procedure can deal with a large number of variables, including a numeric and nonnumeric. The nonnumerical values are converted into numerical value during the processing. The glm's are standardized with a mean value of 0 and standard deviation of 1. The GLM equation with standardized $\beta$s is:

$$Z_Y \quad = \beta_0 + \beta_1 Z_{x1} + \beta_2 Z_{x2} + \cdots + \beta_k Z_{xk} \tag{14.2}$$

*Z* is calculated by dividing the regression coefficient with standard error. If the *z*-value is big, the true regression coefficient is not 0 and the *X*-variable changes.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{14.3}$$

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as taking the square of residuals and summing all observations. The deviance is used to assess the goodness of fit. The higher number of deviance says that there is a bad fit. The response variable is predicted by the null deviance values, by a model that includes only the intercept (grand mean) where as residual with inclusion of independent variables. The data model is pretty only when the null deviance and residual deviance is small. The model is fitted with glm(formula = Order.method.type ~ Retailer.country, family = binomial(link = "logit"), data = meandata). The Deviance Residuals with Order.method.type ~ Retailer.country is shown in Table 14.1.

The final output for a GLM models displays.

- Call
- Residual
- Coefficient
- Dispersion parameter
- Deviance values

The dispersion parameter for binomial family is taken to be 1, the null deviance is 66.925 on 48 degrees of freedom, and residual deviance is 65.550 on 28 degrees of freedom with AIC as 107.55. The Number of Fisher Scoring iterations is 4. The Fisher's scoring algorithm is a method for providing solution to maximum likelihood problems on numerical values. This is derivative of newton's method to perform fit. The fit is performed with glm(formula = Order.method.type ~ Retailer.country + Revenue + Planned.revenue + Product.cost + Quantity, family = binomial(link = "logit"), data = meandata) and is shown in Table 14.2.

**Table 14.1** Deviance residuals with Order.method.type ~ Retailer.country

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| −1.4826 | −1.1774 | 0.9005 | 1.1774 | 1.1774 |

**Table 14.2** Deviance residuals

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| −1.8122 | −1.1774 | 0.4454 | 1.0957 | 1.4579 |

The ANOVA is an analysis of variance which is performed for comparing three or more mean value. It is used to determine if there exist any relationships among variables or any difference between groups in sample data. ANOVA can be used with *t*-tests, Regression, and Chi Square, whether the mean of a variable is less than, greater than, or equal to a specific value. Usually, the known value which is present in the database is a population mean. The Null hypothesis says there is no significant difference between the sample mean and the population mean.

The ANOVA is performed for glm mean model with Chi square test and binomial model link logit. The Analysis of Deviance is shown in Table 14.3. The response attribute used is order method type.

The ANOVA is performed for linear regression model with unit sale price and unitprice as features. The Analysis of Variance is shown in Table 14.4. The response variable is unitsaleprice.

The null hypothesis for ANOVA states that all population means are exactly equal. The significant level of 0.05 shows the mean population or sample used in this work is good. A significance level of 0.05 indicates a 5% risk. This means that a difference exists in the sample mean when there is no actual difference found. This shows the sample means will differ a bit.

The dispersion parameter for binomial family is taken to be 1, the null deviance is 66.925 on 48 degrees of freedom, and residual deviance is 57.901 on 24 degrees of freedom with AIC as 107.9. The Number of Fisher scoring iterations is 17. It is shown that the null variance in the first model and second model is same whereas the residual deviance of second model is decreased with decreased degrees of freedom. The significant reduction in residual deviance shows the improvement of goodness of fit and is shown in Table 14.5. The comparison results are shown in Fig. 14.9.

**Table 14.3** Response-Order.method.type

| Response | Deg of freedom | Deviance | Resid.Df | Resid.Dev | Pr(>Chi) |
|---|---|---|---|---|---|
| NULL | | 48 | 66.925 | | |
| Retailer.country | 20 | 1.3752 | 28 | 65.550 | 0.711 |

**Table 14.4** Response: Unit.sale.price

| Response | Deg of freedom | Sum square | Mean square | F score value Pr(>F) |
|---|---|---|---|---|
| Residuals | 48 | 12.181 | 0.25377 | 0.005 |

**Table 14.5** Comparison of GLM model

| F score | Null variance | Residual deviance | AIC |
|---|---|---|---|
| 4 | 66.295 (48 DF) | 65.550 (28 DF) | 107.55 |
| 17 | 66.925 (48 DF) | 57.901 (24 DF) | 107.9 |

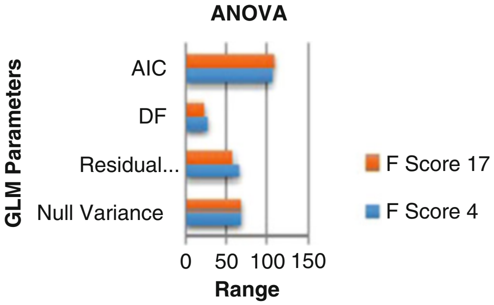


**Fig. 14.9** Comparison of GLM Model

## 14.5 Conclusion

The important outcome of this work is choosing most appropriate missing value handling method. Mean imputation method was proposed to overcome the missing value. The Decision tree classification algorithm was performed. The experimental results show that there is a good improvement in the accuracy of classification algorithm after imputing mean value in the dataset. The goodness of fit of mean imputed value is also analyzed. The future work focuses on implementing other machine learning algorithms with increased results.

## References

1. I. Pratama, A. Erna Permanasari, I. Ardiyanto, R. Indrayani, A review of missing values handling methods on time-series data, in *International Conference on Information Technology Systems and Innovation (ICITSI)* (IEEE, Piscataway, NJ, 2016), INSPEC Accession Number: 16675571
2. M. Moore, J.M. Carpenter, A decision tree approach to modeling the private label apparel consumer. Mark. Intell. Plan. **28**(1), 59–69 (2010)
3. K. Maheswari, P. Packia Amutha Priya, Predicting customer behavior in online shopping using SVM classifier, in *IEEE International Conference on Intelligent Techniques in Control, Optimization, Signal Processing, INCOS'17*, 1 Mar 2018
4. K. Maheswari, P. Packia Amutha Priya, Analysis and implementation of text mining for different documents. Int. J. Scient. Res. Sci. Technol. **3**(5), 109–113 (2017). ISSN: 2395-6011
5. K. Maheswari, P. Packia Amutha Priya, Classification of twitter data set using SVM and KSVM. Int. J. Pure Appl. Math. **118**(7), 675–680 (2018). ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version), Scopus Indexed
6. K. Maheswari, Improving accuracy of sentiment classification analysis in twitter data set using knn. Int. J. Res. Anal. Rev. **5**(1), 422–425 (2018). E ISSN: 2348-1269, Print ISSN: 2349-5138, UGC Approved Journal
7. R. Tinabo, Decision tree technique for customer retention in retail sector, in *International Conference on Integrated Computing Technology INTECH 2011* (Springer, Berlin, 2011), pp. 123-131
8. M. Rafiqul Islam, M. Ahsan Habib, A data mining approach to predict prospective business sectors for lending in retail banking using decision tree. Int. J. Data Min. Knowl. Manag. Process (IJDKP) **5**(2), 13–22 (2015)

9. B.N. Patel, S.G. Prajapati, K.I. Lakhtaria, Efficient classification of data using decision tree. Bonfring Int. J. Data Min. **2**(1), 6–12 (2012)
10. L. Li, X. Zhang, Study of data mining algorithm based on decision tree, in *2010 International Conference on Computer Design and Applications*, 5 Aug 2010, INSPEC Accession Number: 11523965
11. R. Senapati, K. Shaw, S. Mishra, D. Mishra, A novel approach for missing value imputation and classification of microarray dataset. Proc. Eng. **38**, 1067–1071 (2012)
12. R. Houari, A. Bounceur, T. Kechadi, T. Abdelkamel, R. Euler, A new method for estimation of missing data based on sampling methods for data mining, in *Advances in Intelligent Systems and Computing (AISC)*, vol. 225, (Springer, Cham, 2012), pp. 89–100
13. P.R. Houcka, S. Mazumdarb, E. Hartin, Classification of missing values handling method during data mining: review. Sigma Epsilon **21**(2), 49–60 (2017). ISSN: 0853-9103
14. Y.-Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry **27**(2), 130–135 (2015)
15. S. Agarwal, G.N. Pandey, M.D. Tiwari, Data mining in education: data classification and decision tree approach. Int. J. e-Education, e-Business, e-Management and e-Learning **2**(2), 140–144 (2012)
16. C. Kishor Kumar Reddy, B. Vijaya Babu, A survey on issues of decision tree and non-decision tree algorithms. Int. J. Artif. Intell. Appl. Smart Dev. **4**(1), 9–32 (2016)
17. X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, et al., Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37
18. H. Yang, S. Fong, Optimized very fast decision tree with balanced classification accuracy and compact tree size, in *2011 Third International Conference on Data Mining and Intelligent Information Technology Applications (ICMiA)*, 24–26 Oct 2011, pp. 57–64