Comparative Study of Missing Value Imputation Techniques on E-Commerce Product Ratings

Dimple Chehal, Parul Gupta, Payal Gulati, Tanisha Gupta
Department of Computer Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad, India
E-mail: dimplechehal@gmail.com, parulgupta_gem@yahoo.com, gulatipayal@yahoo.co.in, tanishagupta067@gmail.com

Keywords: imputation techniques, missing value, recommender system, sparsity, user ratings

Received: May 5, 2022

Missing data is typical as it adds ambiguity to data interpretation, and missing values in a dataset represent loss of vital information. It is one of the most common data quality concerns, and missing values are typically expressed as NANs, blanks, or other placeholders. Missing values create imbalanced observations, biased estimates and sometimes lead to misleading results. As a result, to deliver an efficient and valid analysis, there arises a need to take the solutions into account appropriately. By filling in the missing values, a complete dataset can be created and the challenge of dealing with complex patterns of missingness can be avoided. In the present study, eight different imputation methods: SimpleImputer, KNN Imputation (KNN), Hot Deck, Linear Regression, MissForest, Random Forest Regression, DataWig, and Multivariate Imputation by Chained Equation (MICE) have been compared. The comparison has been performed on Amazon cell phone dataset based on three parameters: R- Squared Error (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE). Based on the findings KNN had the best outcomes, while DataWig had the worst results for R- Squared error (R^2). In terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE), the Hot Deck imputation approach fared best, whereas MissForest performed worst for Mean Absolute Error (MAE). The Hot Deck imputation approach seems to be of interest and should be investigated further in practice.

Povzetek: Primerjava tehnik imputiranja manjkajoče vrednosti pri ocenah izdelkov e-trgovine

1 Introduction

Missing data occurs frequently in research such as Clinical Trials, Climatology and Medicine as it adds a layer of ambiguity during data interpretation [9], [19], [1], [5]. Nowadays, most databases present a problem of incomplete data. Missing values in a dataset mean loss of important information. These are values that are not present in the data set and are written as NAN's, blanks, or any other placeholders. Missing value creates imbalanced observations, biased estimates and in some cases can direct to misleading results. There can be multiple reasons for the missing value in a dataset such as failure to capture data, incorrect measurements or defective equipment, data corruption, sample mishandling, a low signal-to-noise ratio, measurement inaccuracy, non-response, or a deleted anomalous result [15], [10]. Building a machine learning algorithm with a dataset containing missing values can have a major impact on machine learning models as well as on the outcomes. Missing values can be of both continuous and categorical types. To get more precise results, multiple techniques can be used to fill out missing values.

Many approaches for dealing with missing data have been presented in recent years, and they can be categorized as deletion and imputation. There are three common deletion approaches list wise deletion, pair-wise deletion, and feature deletion. The common approach in list wise or case elimination is to omit the cases with missing values and evaluate the remaining data. Pair-wise deletion, on the other hand, removes data only when the specific data points required to test a hypothesis are missing. The existing values are employed in statistical testing if there is missing data elsewhere in the data set. A pair-wise deletion maintains more information than a list wise deletion since it uses all information observed [11].

Imputation on the other hand is the process of identifying missing values and interchanging them with a substitute value is known as missing value imputation [13], [6]. The method of missing value imputation is depicted in Figure 1. The experiment begins with the selection of a dataset, which is then characterized as incomplete or complete based on the quantity of missing data in the dataset. When a dataset is classified as incomplete, it is split into two parts: complete data and missing data. Imputation methods employ the entire dataset to impute missing values in the dataset. After that, a complete dataset with no missing values is created. The performance of the imputation methods is computed when the whole dataset and experimental dataset are compared using performance measures.

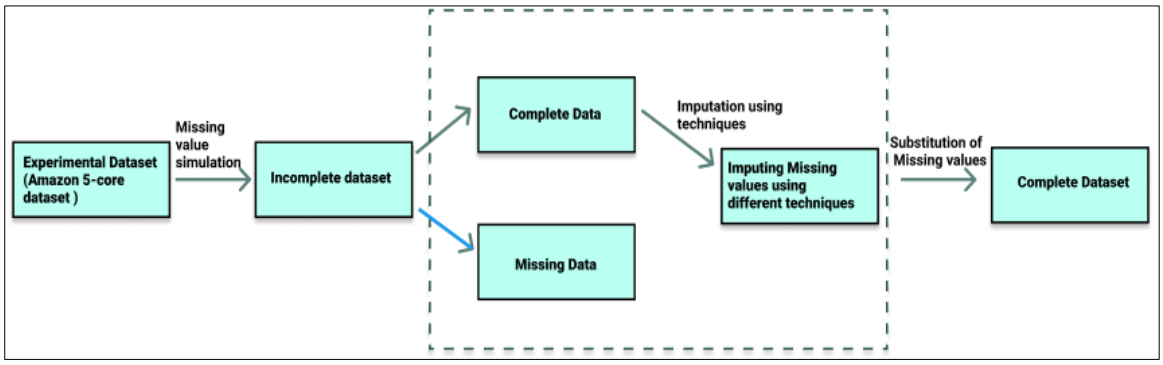

Figure 1: Missing value imputation process.

Single imputation and multiple imputations are two subgroups of the numerous imputation techniques. In single imputation, only one value is present for each missing cell and the value thus generated is used as the original value, although no imputation method can provide the exact value [18], [25]. The workflow for single imputation is depicted in Figure 2. First, the type of missing data is determined, and then single imputation is chosen from the two alternatives of single and multiple imputations, which is further separated into explicit and implicit modeling. The assumptions are explicit because the predicted distribution in explicit modeling is based on a formal statistical model, like multivariate normal. This process employs the mean imputation and regression imputation techniques. Hot Deck imputation, substitution, and cold deck imputation are all part of this procedure.

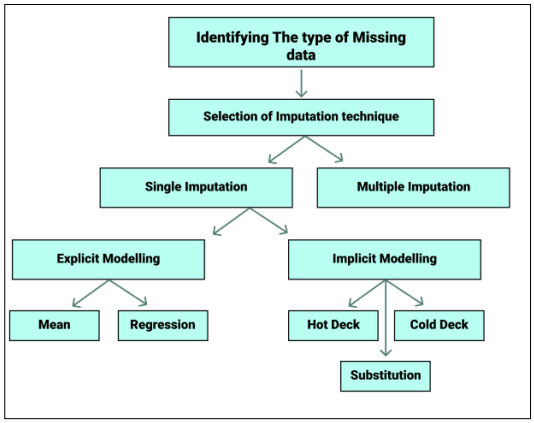

Figure 2: Work flow for single imputation.

In multiple imputations of a missing cell, multiple values are generated to impute the cell. Many complete data sets with various imputed values generate after which each data set is analyzed independently and the results computed. In contrast to single imputations, multiple imputations account for statistical uncertainty in the imputations [21], [7]. The workflow for multiple imputations is depicted in Figure 3. First, the type of missing data is determined, and then multiple imputations are chosen from the two alternatives of single and multiple imputations. Several imputations generates multiple values from separate imputed sets, which are then analyzed after calculating a single value for each missing value, and a single value is chosen from all the values to impute a missing value in the incomplete dataset. As a result, there are three separate phases to the multiple imputation technique:

a. M handles missing data, resulting in M complete data sets.
b. After that, the M full data sets are analyzed.
c. For the final imputation result, the outcomes of all M imputed data sets are pooled.

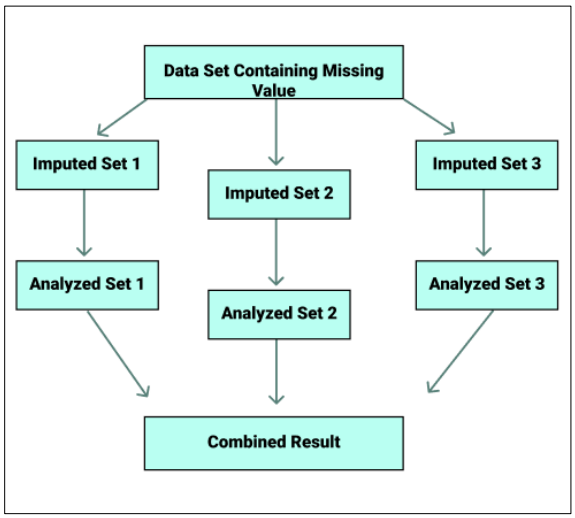

Figure 3: Work flow for multiple imputation.

Existing imputation techniques have been compared using R Squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE) Metrics.

There are three main types of missing values:

a. Missing completely at random (MCAR)
b. Missing at random (MAR)
c. Not missing at random (NMAR)

The relationship between missingness and the values of the variables in the dataset is stated by the missing data mechanism. A dataset Y is stated to be a combination of a variable that is observed and a variable that is missing (Y_{obs} and Y_{mis}, respectively). The first type is known as missing completely at random (MCAR), in which the value itself or any known value is not a determinant of the missing values. Thus, Y_{obs} and Y_{mis} have no effect on the likelihood of a missing value [11], [8], [14]. The second type is Missing at random (MAR) is the polar opposite of MCAR, in which missing

values are dependent on known values or on the value itself. Thus, the probability of a missing value is independent of Y_{mis} or Y_{obs}. MAR and MCAR can be ignored because it is impossible to adjust for the missingness. The final form Not missing at random (NMAR), where the probability of a missing occurrence varies [14].

This study is divided into seven sections. A brief past related work is provided in Section 2. Different missing value patterns are explained in section 3. In section 4, a description of the dataset as well as data analysis is given. The paper's results, as well as the evaluation criteria used, are explained in section 5, and the study's conclusion is shown in section 6.

2 Related work

There are multiple techniques to impute missing value's the first and the oldest one is SimpleImputer in which mean of a single column is computed to fill the missing value or cell with the mean computed of the rest of the cells of that column. SimpleImputer leads to poor imputation because it ignores correlation between different features [14].

Whenever the variables have a non-linear connection, linear regression-based imputation may underperform. The conditional model for imputation is Classification and Regression Trees (CART) [3]. Random forest extensions also have yielded encouraging results [22]. The decision tree-based imputation techniques are non-parametric algorithms that do not forecast the distribution of the data.

K-Nearest Neighbors (K-NN) based imputation is one of the most often used non-parametric techniques. This technique replaces the observed values in dimension d for each missing element with the mean of the K-nearest neighbors' d^{th} dimension [24]. Sequential K-NN is a K-NN extension that begins by imputing missing values from observations with the fewest missing dimensions and then moves on to the next unknown entries while reusing the previously imputed values [12]. Iterative K-NN uses an iterative procedure to re-estimate the estimates and select the closest neighbors based on the previous iteration's estimations.

Single imputation approaches produce a single set of finished data that may be utilized for statistical analysis. Whereas, multiple imputations, impute numerous times (each set may be different), then conduct statistical tests on all sets and combine the results. This strategy can capture the variability in missing data and, as a result, produce potentially more accurate estimates for the wider statistical problem. Multiple imputation approaches, on the other hand, are slower and necessitate pooling of results, which may not be appropriate for some applications.

The process for generating several estimates of missing data varies within the multiple imputation frameworks. A common multiple imputation method, multivariate imputation by chained equations (MICE), generates estimates using predictive mean matching, Bayesian linear regression, logistic regression, and other techniques [4]. Missing data imputation is still a hot topic in research because of its importance. Despite the fact that there are several approaches, many of them have serious shortcomings and their own pros.

In the event of missing values, information management is critical. Planning, organizing, structuring, processing, regulating, assessing, and reporting information operations are all part of the information management cycle. The major goal of information management is to produce and manage data in order to gain better insights; hence in missing value imputation, missing data is discovered using various strategies both of single imputation and multiple imputations in order to gain a better understanding of datasets and compute important and numerically significant conclusions. When managed information is fed into any algorithm, the algorithm's performance improves, ultimately assisting in the resolution of recent technological issues.

3 Missing data patterns and imputation approaches

Missing data patterns explain which values in the dataset are missing and which values should be observed. Univariate, monotone, and non- monotone missing data patterns are the three types of missing data patterns.

a. ***Univariate:*** When only one variable has missing data, the data is classified as univariate missing data pattern. To be classified in Univariate, the missing values should be in one column [17].
b. ***Monotone:*** When data is ordered and the pattern is frequently connected with longitudinal studies where participants drop and never return, it is called Monotone data. This is easier to detect because they are more visible and distinguishable [2].
c. ***Non- Monotone:*** Data is non-monotone when missing values in one variable or column have no effect on the values of other columns or the missing values of other columns [20].

Missing value imputation is the most important part of data analysis since it ensures that the dataset is complete and the results are computed correctly. There are mainly two types of imputation techniques single imputation and multiple imputations. In this experiment, techniques like SimpleImputer, KNN Imputation (KNN), Hot Deck, Linear Regression, MissForest, Random Forest Regression, DataWig, and Multivariate Imputation by Chained Equation (MICE) will be compared and evaluated. Advantages and disadvantages of these techniques have been shown in Table 1.

a. ***Imputation using SimpleImputer:*** SimpleImputer is a scikit-learn class that aids with missing data imputation in datasets used for predictive modeling [16], [23]. It substitutes a placeholder for the NaN values. SimpleImputer employs a variety of strategies to impute values, one of which is the use of mean/median to replace missing values. In this technique, the mean or median of the non-

missing values is computed, and the missing values in the column are imputed using the computed mean or median value. This technique is best applied to numerical values rather than categorical ones. Mean imputation is quick and simple to implement, it preserves the mean of the observed data. This implies if data is Missing completely at random (MCAR), the estimate of mean remains unbiased. However, mean imputation is less accurate than other impute techniques.

b. ***Imputation using KNNImputer***: KNNImputer is a *scikit-learn* python machine learning library that aids in nearest neighbor imputation [16]. In KNN imputation, the distance between data points is measured and the number of contributing neighbors is chosen for each prediction. The number of nearest neighbors used to predict a missing value is usually controlled by the value of K, which has a direct impact on the KNN algorithm's performance. A high K value reduces the impact of random error on variance, but it also increases the risk of missing important small-scale patterns. When selecting an appropriate value of K, it is critical to strike a balance between over fitting and under fitting.

c. ***Hot Deck imputation***: In a sample set with similar values on all other variables, Hot Deck imputation selects one value at random from each individual set of values. This means that all records in the dataset with similar values in other variables are searched, and any one record is selected and utilized to impute the missing values [17].The benefit is that no outliers are created in the dataset as a result of this method.

d. ***Imputation using Linear Regression***: Regression is a two-step procedure in which a regression model is first constructed utilizing all of the available and complete data points. The created model is then used to impute missing data. In linear regression a regression equation is formed in which the best predictors are classed as independent variables, whereas variables with missing data are labeled as dependent variables. The missing values are predicted using a regression equation using independent and dependent variables. Values for the missing variable are inserted in an iterative procedure, and then all cases are utilized to forecast the dependent variable. These steps are repeated until the projected values are almost identical from one step to the next, at which point they converge.

e. ***Imputation using MissForest***: MissForest is a machine learning data imputation method that is based on the random forest algorithm [22]. Firstly the missing data are imputed using median/mode imputation. Then the non-missing values are marked as training rows and missing values are marked as predicted, the training rows are fed into a random forest model used to predict the missing values. The training rows are then fed into a random forest model that predicts missing values. The projected values are then imputed to replace the existing values, resulting in a dataset that is full and free of missing values. To enhance imputation in each iteration, the entire procedure is done numerous times. MissForest is capable of handling numerical, categorical, and mixed data types. MissForest is created with the *missingpy* library.

f. ***Imputation using Random Forest Regression***: The Random Forest is a Meta estimator technique that employs averaging to increase predicted accuracy and control over-fitting by fitting several classification decision trees on various sub-samples of the dataset. Random forest regression is a supervised learning approach for regression that uses the ensemble learning method. The ensemble learning method combines predictions from several machine learning algorithms to get a more accurate forecast than a single model. For regression problems, the mean or average forecast of the individual trees is computed known as aggregation. Instead of depending on individual decision trees, the main idea is to aggregate numerous decision trees to determine the final outcome. As a fundamental learning model, Random Forest uses several decision trees. Row and feature sampling are done at random from the dataset, resulting in sample datasets for each model this process is known as bootstrap.

g. ***Imputation using Deep Learning (DataWig)***:DataWig is a machine learning package that employs Deep Neural Networks to impute missing values in a dataset [2]. DataWig combines deep learning feature extraction with automatic hyper parameter tuning. This approach applies to both categorical and non-numerical data. DataWig first determines the type of each column. The column is then translated to a numerical representation. DataWig can be used to train on both the CPU and the GPU. DataWig typically works on a single column at a time, with the target column holding information about the imputing column supplied ahead of time.

h. ***Imputation using Multivariate Imputation by Chained Equation (MICE)***: In multiple imputations, many imputations are created for each missing value. It means filling the missing values multiple times and creating multiple complete datasets. One well-known algorithm for multiple imputations is Multiple Imputation by Chained Equation (MICE). MICE works under the assumption that missing data is Missing at random (MAR) or Missing completely at random (MCAR). Implementing MICE when data is not MAR could result in biased estimates. MICE is very flexible

technique and can handle multiple variables and complexities of varying types at a time. It employs a divide-and-conquer strategy to impute missing values in dataset variables, focusing on one variable at a time. Once the emphasis is placed on that variable, it uses all of the other variables in the data set to forecast missingness in that variable. A regression model, the form of which is dictated by the nature of the focal variable, is used to make the prediction.

Table 1: Advantages and disadvantages of imputation techniques

S. No	Method	Advantages	Disadvantages
1.	SimpleImputer	1. It's a simple and quick procedure. 2. It's suitable for small numerical datasets.	1. Correlation between features is not taken into account. 2. Not extremely precise.
2.	KNNImputer	1. Better than SimpleImputer in terms of accuracy	1. KNN operates by memorizing the entire training dataset 2. Sensitive to outliers
3.	Hot Deck imputation	1. Because of residuals, the imputed data will have the same distribution shape as the actual data. 2. It's good for categorical data.	1. It's not good for small sample sizes.
4.	Linear Regression	1. For numeric data, this strategy is more effective.	1. If the prediction power is poor, this approach will perform poorly.
5.	MissForest	1. The looping over missing data point's process is repeated numerous times, with each iteration improving on improved data. 2. It can be used with both numerical and category data. 3. There is no need for preprocessing.	1. Time consuming because the number of iterations is dependent on the size of the dataset. 2. Expensive to operate MissForest
6.	Random Forest Regression	1. Outlier resistant. 2. Does a good job with non-linear data. 3. Less chance of over fitting. 4. Performs well on a huge dataset.	1. Slow and steady training. 2. Linear approaches with a lot of sparse features aren't recommended.
7.	Deep Learning (DataWig)	1. It works with categorical data. 2. Supports both CPUs and GPUs	1. Slow when dealing with large datasets 2. Imputation of a single column.
8.	Multivariate Imputation by Chained Equation (MICE):	1. Unbiased estimates, which are more reliable than ad hoc responses to missing data	1. MICE works under the assumption that missing data is Missing at random (MAR) or Missing completely at random (MCAR)

4 Experiments on rating predictions

This section details the dataset used and its corresponding analysis.

4.1 Dataset description

In this study, the publicly accessible dataset from Amazon of cell phone and accessories has been used. In the 5-core dataset, all users and items have at least five reviews. It consists of 1048570 rows and 12 columns. The 12 columns are overall (rating of product), Verified (for verified product by Amazon), ReviewTime (time of review submission), ReviewerID (ReviewerID of each reviewer), Asin (product ID), Style (sparse value pertaining to product's color), ReviewerName (name of the reviewer), ReviewText (review text), Summary (review summary), UnixReviewTime (review time (UNIX time)), Vote (total number of votes earned by a product), Image (product image link).

The primary columns to pay attention are verified, vote and rating. Then the dataset is preprocessed to ensure that every product has a vote value because the data is massive and sparse in the vote column. The dataset was reduced to 90714 rows and 12 columns after preprocessing.

4.2 Data analysis

The principle of analysis is depicted in Figure 4. Initially, there were no missing values in the dataset. As a result, missing values of about 4\% were created in the original dataset (Amazon 5-core) based on the MCAR model in the overall column, and imputation was performed using several strategies. These missing values were simulated and imputed using the eight techniques and three evaluation criteria (R-squared error, MAE, and MSE). R-squared, a statistical measure represents the degree of goodness of fit of a regression model. The best r-square value is 1.

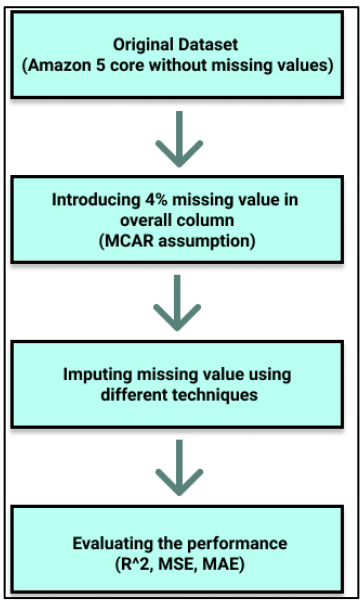

Figure 4: Principal of analysis

5 Results and discussion

Missing values were imputed using eight distinct imputation approaches. With the use of the vote and verified columns, all of the strategies effectively imputed the missing values that were present in the Overall column. The three-assessment metrics were used to measure the performance of the techniques R^2, MSE and MAE, Table 2 compares all the eight approaches based on these assessment metrics.

a. ***R-squared:*** The closer the r-squared value is to 1, the better the model fits. When the fitted models are worse than the average fitted model, the R-Squared value can be negative. The R-squared is determined by dividing the sum of squares of residuals from the regression model (SS_{RES}) by the total sum of squares of errors from the average model (SS_{TOT}), then subtracting 1. The R-squared is mathematically defined by the equation 1:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_j - \hat{y}_j)^2}{\sum_i (y_j - \bar{y}_j)^2} \quad (1)$$

Results for the R-squared (R^2) metrics: R^2 usually has a range of 0 to 1. Figure 5 shows graph for R^2. All eight approaches yielded a value ranging from -0.5 to 1 for R^2. R^2 values that are negative indicate that the fitted models are worse than the average fitted model. KNN with value 0.9742 is the approach that produced the best R^2 value. When computing the missing value in KNN, the K is set to 4, implying that the value for a missing point is computed using four nearest neighbors. DataWig, on the other side with an R^2 of -0.5311, had the poorest performance. SimpleImputer, Hot Deck, MICE, and Random Forest Regression all received positive results, with values of 0.9744, 1.0, 0.9929, 0.97443, and 0.9745, respectively. Linear Regression and MissForest, on the other hand, calculated negative R^2 values of -0.4356 and -0.0259, respectively.

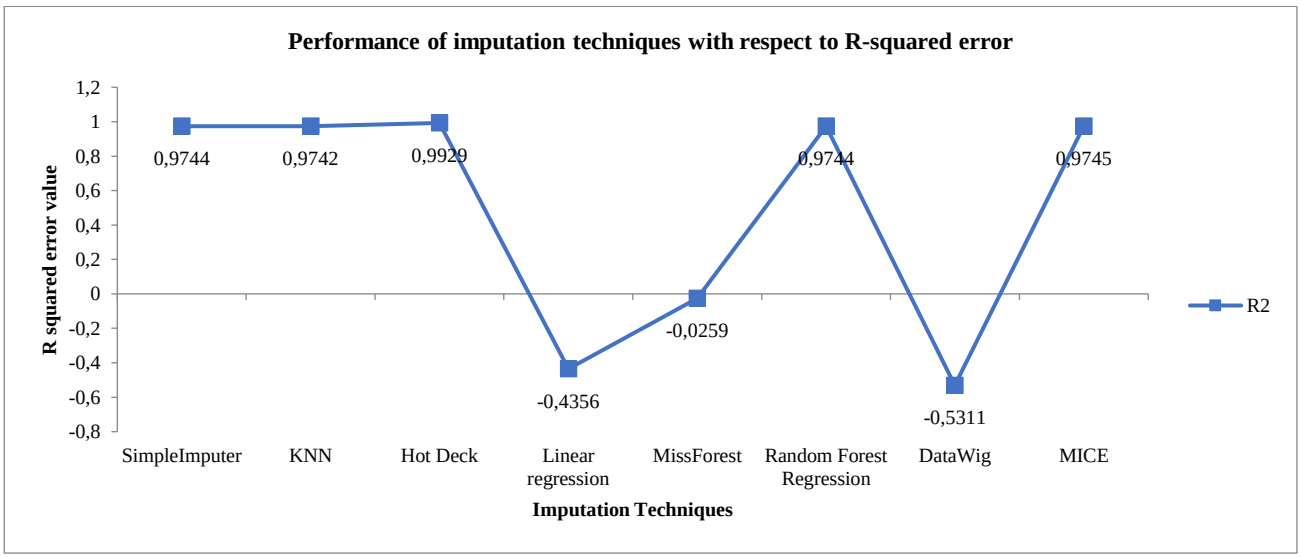

Figure 5: Graphical representation of comparison of imputation techniques with respect to R-squared error.

b. Mean Squared Error: The Mean Squared Error (MSE) is one of the most basic and often used loss functions. To calculate the MSE, take the difference between model's predictions and the ground truth, square it, and average it over the whole dataset. The value of MSE can never be negative because errors are always squared. The amount of samples tested is denoted by N. The advantage with MSE is that it is useful for ensuring that our trained model does not contain any outlier predictions with significant mistakes, as the squaring element of the function gives these errors more weight. The MSE is mathematically defined by the equation 2:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \qquad (2)$$

Result for Mean squared Error (MSE) metrics: The Mean Squared Error ranges from 0 to infinity. Figure 6 shows graph for MSE. The value point for MissForest is out of the range when compared to the other points; hence it isn't depicted in this graph. The MSE regression is the most widely used regression for loss functions. Because the real and predicted values are so near, the lower the MSE value, the higher the predicted values accuracy. MissForest (1207.2801) is the strategy that produced the highest MSE while Hot Deck (0.0145) produced the lowest value. MSEs are smaller than 1 for SimpleImputer (0.0514), KNN (0.0529), Random Forest regression (0.0515), and MICE (0.0513) and Linear Regression (1.2888) and DataWig (1.3746) have MSEs more than 1.

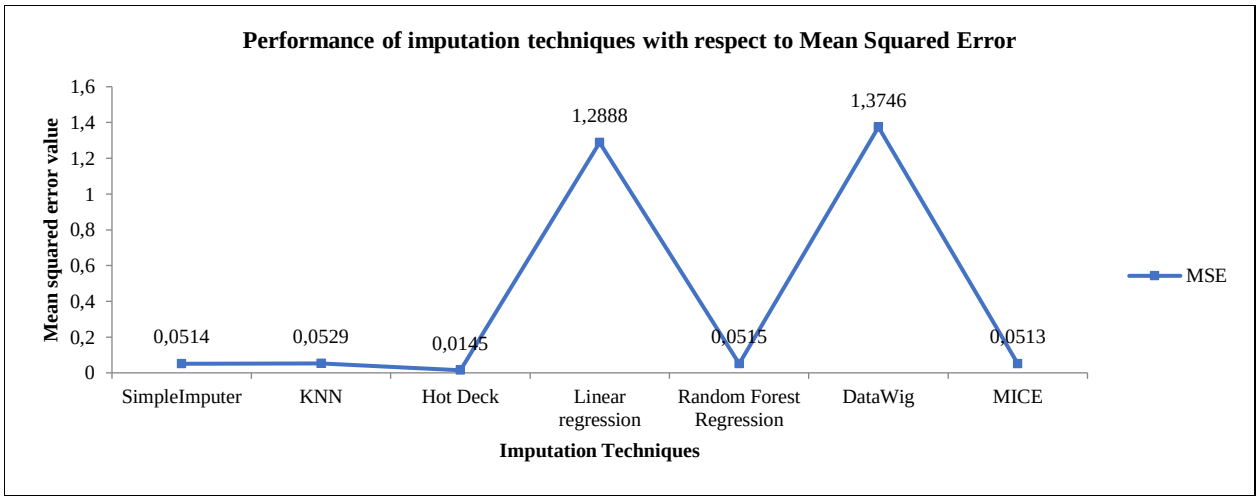

Figure 6: Graphical representation of comparison of imputation techniques with respect to MSE.

c. Mean Absolute Error: The difference between the model's predictions and the ground truth is used while computing the Mean Absolute Error (MAE) and the absolute value is applied to the difference and averaged throughout the entire dataset. The MAE advantage compensates for the MSE disadvantage directly. Because the absolute value is considered, all errors will be weighted on the same linear scale. As a result, unlike the MSE, the loss function will not place an excessive emphasis on outliers and will provide a general and consistent evaluation of how well our model is performing. The MAE is mathematically defined by the equation 3:

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j| \qquad (3)$$

Results for Mean Absolute Error (MAE) metrics: Mean absolute error ranges from 0 to infinity. Figure 7 shows graph for MAE. Initially, the MAE error is calculated in phases. By subtracting the predicting value from the actual value, the prediction error is calculated. Then, for each imputation, the prediction error is calculated and transformed to positive values. It is determined what the mean of all absolute errors is. The best MAE results were achieved by Hot Deck (0.0052), while the poorest MAE results was achieved by MissForest (7.6032). Other techniques produced results ranging from 0 to 1 such as MICE (0.0410), SimpleImputer (0.0411), KNN (0.0245), Linear Regression (1.0319), Random Forest Regression (0.0410) and DataWig (1.0768). The result of measuring the difference between any two continuous variables is generally referred to as Mean Absolute Error.

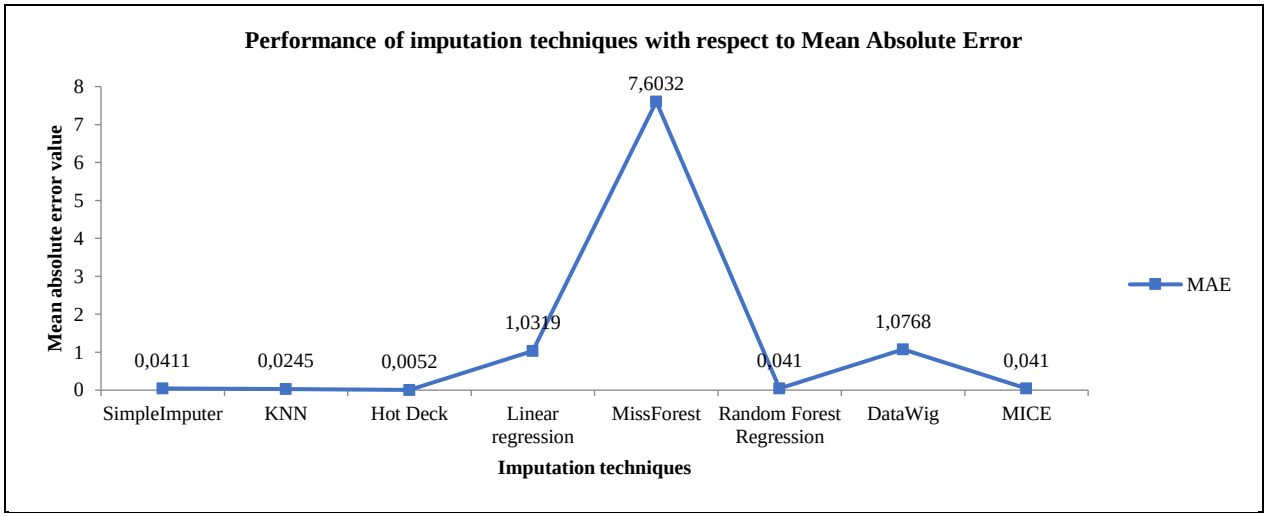

Figure 7: Graphical representation of comparison of imputation techniques with respect to MAE.

As shown in Table 2 Hot Deck imputation technique is the best technique that provides the most promising outcomes and should be considered further, while MissForest produced the worst results. All of the other strategies produced outcomes that might be improved over time by making simple adjustments.

Table 2: Performance comparison of imputation techniques

Techniques	R^2	MSE	MAE
SimpleImputer	0.9744	0.0514	0.0411
KNN	0.9742	0.0529	0.0245
Hot Deck	0.9929	0.0145	0.0052
Linear regression	-0.4356	1.2888	1.0319
MissForest	-0.0259	1207.2801	7.6032
Random Forest Regression	0.9744	0.0515	0.0410
DataWig	-0.5311	1.3746	1.0768
MICE	0.9745	0.0513	0.0410

6 Conclusion

When a value in a dataset goes missing, important information is lost. To avoid this, missing values are imputed. The term "imputing values" refers to the statistical computation of a value for a missing value based on surrounding values or values from the same column. In data analysis, post imputation is significant because it ensures that the dataset is complete and that the findings are computed and arranged accurately. Eight techniques have been explored in this experiment to compute missing values for the Amazon dataset. Only the three columns (Overall, Verified, and Vote) have been utilized to conduct the experiment. Overall column contains missing values, and hence is the most essential column. After imputing the missing values accurately, the outcomes have been evaluated using three evaluation parameters-R^2, MAE and MSE. Hot Deck Imputation technique has surpassed all other techniques in terms of imputation results. The performance metrics for Hot Deck are within the range; however, MissForest's values are outside the range, making it the lowest performing technique.

References

[1] Afrifa-Yamoah, E. et al. 2020. Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*. 27, 1 (2020), 1–18. DOI:https://doi.org/10.1002/met.1873.

[2] Bießmann, F. et al. 2019. DataWig: Missing value imputation for tables. *Journal of Machine Learning Research*. 20, (2019), 1–6.

[3] Burgette, L.F. and Reiter, J.P. 2010. Multiple Imputation for Missing Data via Sequential Regression Trees. *American Journal of Epidemiology*. 172, 9 (Nov. 2010), 1070–1076. DOI:https://doi.org/10.1093/AJE/KWQ260.

[4] Chhabra, G. et al. 2019. A review on missing data value estimation using imputation algorithm. *Journal of Advanced Research in Dynamical and Control Systems*. 11, 7 Special Issue (2019), 312–318.

[5] Cismondi, F. et al. 2013. Missing data in medical databases: Impute, delete or classify? *Artificial Intelligence in Medicine*. 58, 1 (2013), 63–72. DOI:https://doi.org/10.1016/j.artmed.2013.01.003.

[6] Ghazanfar, M.A. and Prugel-Bennett, A. 2013. The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved SVD-based recommendations. *Informatica (Slovenia)*. 37, 1 (2013), 61–92.

[7] Graham, J.W. et al. 2003. Methods for Handling Missing Data. *Handbook of Psychology*. (2003).

DOI:https://doi.org/10.1002/0471264385.wei0204.

[8] Heitjan, D.F. and Basu, S. 1996. Distinguishing "missing at random" and "missing completely at random." *American Statistician*. 50, 3 (1996), 207–213. DOI:https://doi.org/10.1080/00031305.1996.10474381.

[9] Jakobsen, J.C. et al. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology*. 17, 1 (2017), 1–10. DOI:https://doi.org/10.1186/s12874-017-0442-1.

[10] Kaiser, J. 2014. Dealing with Missing Values in Data. *Journal of Systems Integration*. (2014), 42–51. DOI:https://doi.org/10.20470/jsi.v5i1.178.

[11] Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*. 64, 5 (2013), 402. DOI:https://doi.org/10.4097/kjae.2013.64.5.402.

[12] Kim, K.Y. et al. 2004. Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*. 5, 1 (Oct. 2004), 1–9. DOI:https://doi.org/10.1186/1471-2105-5-160/FIGURES/3.

[13] Lin, W.C. and Tsai, C.F. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*. 53, 2 (2020), 1487–1509. DOI:https://doi.org/10.1007/s10462-019-09709-4.

[14] Little, R.J.A. and Rubin, D.B. 2014. Statistical analysis with missing data. *Statistical Analysis with Missing Data*. (Jan. 2014), 1–381. DOI:https://doi.org/10.1002/9781119013563.

[15] Mandel J, S.P. 2015. A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*. 06, 01 (2015), 1–6. DOI:https://doi.org/10.4172/2155-6180.1000224.

[16] McAuley, J. et al. 2015. Image-based recommendations on styles and substitutes. *SIGIR 2015 - Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. (2015), 43–52. DOI:https://doi.org/10.1145/2766462.2767755.

[17] Myers, T.A. 2011. Goodbye, Listwise Deletion: Presenting Hot Deck Imputation as an Easy and Effective Tool for Handling Missing Data. *Communication Methods and Measures*. 5, 4 (2011), 297–310. DOI:https://doi.org/10.1080/19312458.2011.624490.

[18] Plaia, A. and Bondì, A.L. 2006. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*. 40, 38 (2006), 7316–7330. DOI:https://doi.org/10.1016/j.atmosenv.2006.06.040.

[19] Ropper, A.H. et al. 2012. Hyperosmolar Therapy for Raised Intracranial Pressure. *New England Journal of Medicine*. 367, 26 (2012), 2554–2557. DOI:https://doi.org/10.1056/nejmc1212351.

[20] Schuetz, C.G. 2008. Using neuroimaging to predict relapse to smoking: role of possible moderators and mediators. *International journal of methods in psychiatric research*. 17 Suppl 1, 1 (2008), S78–S82. DOI:https://doi.org/10.1002/mpr.

[21] Sinharay, S. et al. 2001. The use of multiple imputation for the analysis of missing data. *Psychological Methods*. 6, 3 (2001), 317–329. DOI:https://doi.org/10.1037/1082-989x.6.4.317.

[22] Stekhoven, D.J. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 28, 1 (Jan. 2012), 112–118. DOI:https://doi.org/10.1093/BIOINFORMATICS/BTR597.

[23] Tan, Y. et al. 2018. Probability matrix decomposition based collaborative filtering recommendation algorithm. *Informatica (Slovenia)*. 42, 2 (2018), 265–271.

[24] Troyanskaya, O. et al. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 17, 6 (Jun. 2001), 520–525. DOI:https://doi.org/10.1093/BIOINFORMATICS/17.6.520.

[25] Zhang, Z. 2016. Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*. 4, 1 (2016). DOI:https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.