Iranian Journal of Finance

Print ISSN 2676-6337
Online ISSN 2676-6345

Comparative Analysis of Missing Values Imputation Methods: A Case Study in Financial Series (S&P500 and Bitcoin Value Data Sets)

Mahdi Goldani*

*Corresponding Author, Assistant Prof., Department of Finance, Hakim Sabzevari University, Sabzevar, Iran. (Email: mahgoldani@gmail.com)

Iranian Journal of Finance, 2024, Vol. 8, No.1, pp. 47-70.
Publisher: Iran Finance Association
doi: https://doi.org/ 10.30699/IJF.2024.414027.1427
Article Type: Original Article

Received: August 25, 2023
Received in revised form: February 12, 2024
Accepted: March 01, 2024
Published online: March 08, 2024

Abstract

The accurate imputation of missing values in time series data is paramount for maintaining the integrity and reliability of analyses and predictions. This article investigates the effica-cy of various missing values imputation methods, encom-passing well-known machine learning and statistical tech-niques. Moreover, for a better understanding, they imple-mented two financial data time series: S&P 500 and Bitcoin markets spanning from 2016 to 2023 on a daily frequency. Initially utilizing complete datasets, controlled missingness was introduced by randomly removing 45 data points. Then, these methods applied multiple imputation strategies for estimating and substituting these missing values. Experi-mental evaluation yielded insightful findings regarding the performance of the different methods. The examined ma-chine learning methods, including k-Nearest Neighbors (k-NN), Random Forest, Deep

Learning, and Decision Trees, consistently outperformed their statistical counterparts, such as Mean Imputation, Regression Imputation, Hot-Deck Imputation, and Expectation-Maximization Imputation. Nota-bly, Random Forest emerged as the most effective method, showcasing superior performance in terms of accuracy and robustness. Conversely, the Mean Imputation method exhibited com-paratively inferior outcomes, suggesting its limited suitabil-ity for financial time series data. This research contributes to the ongoing discourse on data integrity within finance ana-lytics and serves as a comprehensive guide for practitioners seeking optimal missing values imputation methods. The empirical evidence provided herein advances the under-standing of imputation techniques' relative performance and their application in financial data, facilitating enhanced de-cision-making processes and yielding more reliable predic-tions.

Keywords: Missing Values Imputation, Machine Learning, Statistical Methods, Finance Data, S&P 500, Bitcoin, Time Series Analysis.

Introduction

The modern era of big data has provided abundant opportunities for research and innovation. The availability of vast amounts of data has empowered university researchers to develop and test theories with significant scientific and societal impact. However, despite these advantages, there are also significant challenges. One persistent problem that every data analyst faces is managing missing values. Missing values refer to situations where meaningful data values are either unobserved or hidden, posing obstacles to data analysis (Little & Rubin, 2019).

In some cases, university researchers primarily dealt with missing values by deleting observations with incomplete information (known as listwise deletion or complete case analysis) or editing the data (e.g., replacing missing values with the mean of the respective variable or even with zero). However, handling missing values can lead to inference problems, where incorrect conclusions are drawn from the analysis. Missing values are a recurring issue caused by various reasons, including technical problems, data non-observability, user privacy concerns, human errors, and more. Missing values can occur in tabular and time series data (Sidek et al., 2016). These missing values introduce a significant amount of uncertainty in classification tasks. Therefore, it is essential to identify and carefully manage these missing values. Improper management of missing values can lead to unusual effects, such as increased classification time and high incorrect rates (Moeinol et al., 2022).

Missing data imputation is a common problem in financial data analysis. The absence of crucial financial data points can lead to biased insights, impair investment decisions, hinder credit assessments, and distort risk management models. Market analyses, financial reporting, and regulatory compliance are likewise vulnerable to inaccurate conclusions due to missing data. Moreover, operational inefficiencies, increased portfolio volatility, and diminished fraud detection capabilities can arise, necessitating meticulous data collection, imputation techniques, and sensitivity analysis to address these challenges and ensure accurate financial assessments and decision-making processes.

Rubin (1976), in his seminal paper on missing data for the first time, introduces the two types of missing values: Missing at random (MAR) and missing completely at random (MCAR). MCAR data exhibit random distribution across the variable and lack any relationship with other variables. This means that the missing data are unrelated to any of the other variables in the dataset. MAR data are not missing completely at random (MCAR) but at random conditional on the observed data. This means that the probability of a missing value is related to the observed data, not the unobserved values. Missing random-at-not (MNAR) data was introduced in 1987 by little and Rubin. Data are missing, not at random, if the probability of a value being missing is related to the unobserved values. This means that the missing data are related to the missing values, not just to the observed values. MNAR data is the most challenging type of missing data to deal with. It is not possible to use imputation methods to deal with MNAR data, and it is necessary to use more sophisticated methods, such as model-based methods or inverse probability weighting.

Several different methods can be used to deal with missing data. The best method to use will depend on the specific situation. Some standard methods include:

- Imputation: This involves replacing missing data with estimated values. Various imputation methods are available, each with its advantages and disadvantages.
- Deletion: This involves deleting cases with missing data. This can be a good option if the amount of missing data is small or if the missing data is not randomly distributed.
- Modeling: This involves using statistical models to account for missing data. This can be a good option if the amount of missing data is large or if the missing data is not randomly distributed.

As per findings by Strike et al. (2001) and Raymond and Roberts (1987), when dealing with datasets that have a minimal amount of missing data, such as a missing rate of less than 10% or 15% for the entire dataset, it is generally acceptable to straightforwardly eliminate the missing data. This removal typically has negligible impacts on the ultimate mining or analytical outcomes. However, when the missing rate surpasses 15%, it becomes imperative to approach the handling of these missing data with careful consideration. (Acuna & Rodriguez, 2004; Lin and Tsai, 2021). Missing value imputation can help address the missing data problem by replacing the missing values with estimated values. This can contribute to enhancing the precision and dependability of the outcomes, and it can also help to reduce the bias in the results. Several different methods can be used for missing value imputation. Some standard methods include Mean, Mode, Random, and Multiple imputation.

The best method for missing value imputation will depend on the specific data set and the type of missing data. However, in general, multiple imputation is considered the most robust method. Missing value imputation is a crucial technique that enhances the precision and trustworthiness of study outcomes. However, it is essential to note that there are better solutions than missing value imputation. It is still possible that the imputed values will be biased, and it is essential to be aware of the limitations of missing value imputation.

The primary objective of this research is to investigate the various techniques for imputing missing values using machine learning and statistical methods. The study aims to contribute to understanding the effectiveness and applicability of these methods in handling missing data. To achieve this goal, the research adopts a two-fold approach. Firstly, a comprehensive review of articles published in recent years is conducted, allowing for identifying and selecting the most recurrently employed imputation techniques. Subsequently, the research employs a real-world dataset encompassing the daily data of S&P 500 companies and Bitcoin from 2016 to 2023. By applying the selected imputation methods to this dataset, the study aims to empirically evaluate their performance in addressing missing values across a diverse range of financial data. Through this empirical analysis, the research sheds light on the comparative advantages and limitations of different imputation strategies, thereby enhancing the knowledge base surrounding effective missing data handling in the context of financial datasets.

Literature Review

Types of missing data:

Rubin and Little (1998) classified missing data into three distinct categories: Missing Completely at Random (MCAR), missing at Random (MAR), and Missing Not at Random (MNAR). Missing completely at Random (MCAR) means no relationship between the missing and observed values. The missingness of the data is completely random and unrelated to any other variables or factors. Missing at Random (MAR) implies a systematic association between missing and observed values. It means that the missingness can be predicted based on other variables in the dataset. Missing not at Random (MNAR) refers to missing data that is not random and cannot be explained by the observed variables. This type of missingness occurs when the missing data is related to unobserved factors or reasons that need to be explicitly recorded or known.

In other words, missing data that is not random has been removed or is unavailable. However, the specific reasons for the missingness, such as the time order, spatial location, or the factors causing the data to be missing, are unknown or unrecorded.

Approaches for handling missing data

In the research literature, there are different approaches to handling missing data. Two fundamental approaches for dealing with missing data are deletion and imputation. Below, each of these approaches is explained:

1. Deletion: Deletion, a method for handling missing data, involves the removal of cases or variables with missing information from the analysis. It is a straightforward approach requiring no imputation or estimation of missing values. However, Little and Rubin (2019) have highlighted certain drawbacks associated with deletion, mainly when the missing data is not randomly distributed. Deletion can introduce bias into the analysis. Two primary ways to perform deletion are pairwise and listwise (Dantan et al., 2008). Here are two common types of deletion methods:

a. Listwise Deletion (Complete-Case Analysis): Listwise deletion entails excluding cases with missing data on any variable in the analysis. This results in a dataset consisting only of complete cases ready for analysis. However, it may reduce sample size and potential bias, primarily if missingness is related to the outcome or variables of interest. Listwise deletion frequently serves as the default choice in numerous statistical analyses. Nevertheless, it is suitable only when there are relatively few missing values, as it can introduce bias if

missing data is not entirely random (MCAR), which is rarely the practice case. Furthermore, listwise deletion risks the loss of critical information associated with missing values, ultimately yielding biased parameter estimates (Abidin et al., 2018).

b. Pairwise Deletion: Pairwise deletion, in contrast, retains cases with missing data on specific variables and incorporates them into the analysis for other variables where data are available. This approach maximizes the utilization of available data but can result in varying sample sizes for different analyses. Kang (2013) argued that pairwise deletion allows testing specific assumptions in the presence of missing values, adapting statistical testing to the observed data. One disadvantage of pairwise deletion is that it may produce standard errors that are either underestimated or overestimated (March 1988).2. Imputation: Data imputation replaces missing data with substituted values (Kang, 2013). Imputation involves estimating or filling in missing values based on available information. There are various imputation methods available:

a. Mean imputation: Missing values are replaced with the variable's mean value. While it is simple to implement, it can lead to underestimating variance and distorting relationships.

b. Last observation carried forward (LOCF): Missing values are imputed with the last observed value for the case. This approach presupposes that the data are missing at random and that the last observed value is a reasonable estimate for the missing value.

c. Multiple imputation: Multiple imputation generates multiple plausible values to replace missing data based on statistical models. It accounts for imputation uncertainty and preserves the dataset's variability.

These approaches have advantages and limitations, and the choice of approach depends on the nature of the missing data, the research context, and the assumptions made about the messiness mechanism.

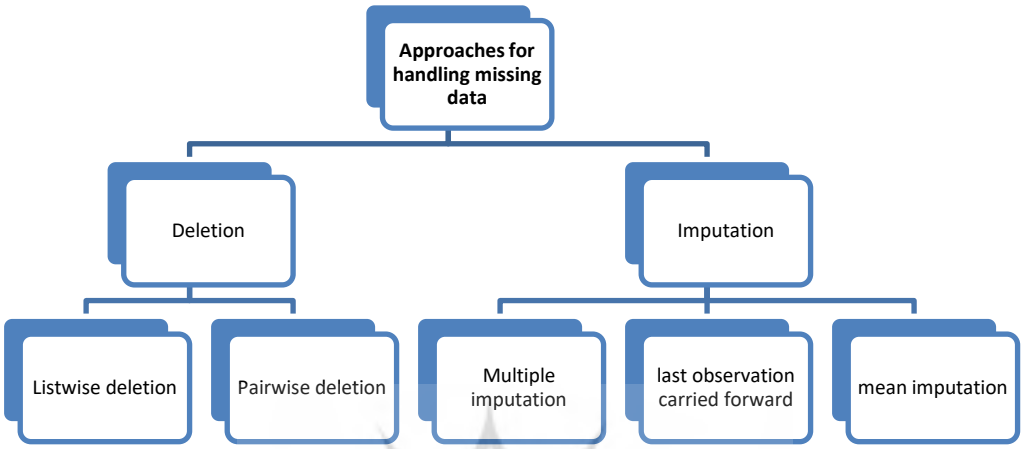


Figure 1. Approaches for handling missing data

Generally, various imputation techniques can be categorized into two groups: Statistical methods and machine learning methods. (wang et al., 2019). It is important to note that the choice of the imputation method depends on various factors, including the missing data pattern, the distribution of the data, the assumptions made about the missingness mechanism, and the research context. Researchers should carefully consider these factors and select an appropriate imputation method accordingly.

Statistical imputation methods

Table 1 provides an overview of the frequently employed statistical techniques utilized in addressing missing data imputation in research studies. Among these prominent methods, mean imputation is the most straightforward technique for numerical attributes, while mode imputation is the equivalent for categorical attributes. In the mean imputation approach, missing values are replaced with the average value of the respective attribute across all available data points. Conversely, in the mode imputation method, the most frequently occurring value in the observed data is used to fill in missing values for categorical attributes.

Table 1. Statistical imputation methods

Methods	Definition	source
Mean and Mode Imputation	Mean and mode imputation is a simple and commonly used method for filling in missing data in a dataset. In this method, the mean or mode of the available values within the same feature (column) is used to fill in a missing value.	Chen et al. (2021).
Regression	In this method, the missing value is imputed based on the other data points using a regression algorithm such as linear regression or logistic regression.	Silva Ramírez et al. (2015)
HOT DECK	The main idea behind this method is to fill in the missing value for each record by using the value from a similar record with the variable present. The imputation with this method occurs in two steps: the first step is clustering, where records are divided into separate homogeneous clusters. In the second step, the complete records within its cluster are used to fill in the missing values for each incomplete record.	Lakshminarayan et al. (1999).
Expectation Maximization	The Expectation-Maximization (E.M.) algorithm is widely employed for managing missing data. It operates through an iterative process to estimate the maximum likelihood parameters of a statistical model, even in scenarios where data points are missing. The EM algorithm demonstrates its effectiveness, especially when the missing data mechanism adheres to the principles of Missing Completely at Random (MCAR) or Missing at Random (MAR).	Little and Rubin(2019)

Imputation methods based on machine learning are intricate processes that typically entail constructing predictive models to estimate values that can be used to replace missing data points. These techniques rely on the information present in the dataset to make these estimations. When the observed data contains valuable information for predicting the missing values, machine learning-based imputation methods can harness this information to achieve high accuracy. There are various machine learning methods for imputing missing values, and here are some commonly employed ones:

Table 2. Machine learning-based imputation methods

Method	Definition	Advantages	Disadvantages	Source
K-Nearest Neighbors (KNN)	In this method, the nearest observed data points to the missing value are found, and the value is imputed based on the values of the neighboring data points.	Simple to implement	Can be sensitive to the choice of K	Hung et al. (2016, 2017)
Decision Tree	This method imputes the missing value using a decision tree algorithm.	Simple to implement	Can be less accurate than regression	Nishat and Ravi (2016)
Random Forest	In this method, several decision trees are constructed using a random forest algorithm, and the missing value is imputed based on these trees.	Can be more accurate than decision trees	Can be more complex to implement	Khia et al. (2017)
Deep Learning	In this method, deep neural networks and algorithms such as autoencoders are used to impute the value of the missing value.	Can be very accurate	Can be very complex to implement	Talmallo et al. (2021)

K-Nearest Neighbors (KNN): K-Nearest Neighbors, abbreviated as KNN, represents a machine learning algorithm widely adopted for imputing missing data due to its robustness and simplicity. KNN imputation works by finding the K nearest neighbors to a data point with missing values and using their values to impute the missing values. The choice of K, the number of neighbors to consider, is an important parameter that can affect the imputation results. One advantage of KNN imputation is that it addresses numerical and categorical data, making it versatile for various datasets. Additionally, KNN imputation needs to make stronger assumptions about the underlying data distribution, which can be beneficial when dealing with complex and diverse datasets (Ismail et al., 2022). However, it is essential to note that KNN imputation has some limitations. A drawback is its potential for high computational costs, especially for large datasets, as it requires calculating the distances between data points. Another limitation is that KNN imputation may perform poorly when the dataset has a high dimensionality or strong correlations between features (zhang et al., 2022).

Decision Tree

One method for addressing missing values is decision tree-based imputation. This approach employs a decision tree algorithm to partition a dataset into distinct segments horizontally. Subsequently, an expectation-maximization (E.M.) algorithm is applied to each segment to impute the missing values within that particular segment (Rahman & Islam, 2014). The decision tree algorithm used can include C4.5 or similar alternatives. In cases where all numerical attribute values within a record are missing, they are substituted with the mean values of attributes from records within the same segment as determined by the decision tree.

Decision tree-based imputation has proven effective in filling missing values and can reduce computational time complexity compared to alternative techniques (Rahman & Islam, 2011). However, it is essential to acknowledge that the performance of decision tree-based imputation may suffer when strong correlations exist between features (Saha et al., 2019).

Random forest (R.F.)

Random Forest (R.F.) algorithms for handling missing data present an appealing approach to imputation. They possess valuable qualities, such as the ability to manage mixed types of missing data, adapt to interactions and nonlinearity in the data, and the capability to scale effectively in big data scenarios. Three primary strategies have been applied for R.F. missing data imputation:

1. **Preimputation Strategy**: In this method, the data is initially preimputed. Then, a forest is grown, and the original missing values are updated based on the proximity of the data. This process can be iterated to refine the results.

2. **Simultaneous Imputation Strategy**: The random forest is constructed here while imputing the missing data. Iterations are performed to enhance the imputed values.

3. **Variable-Specific Imputation Strategy**: In this approach, the data is first preimputed, and then a forest is grown for each variable with missing values. The missing values are predicted using these variable-specific forests, and iterations can be carried out to improve the results (Tang & Ishwaran, 2017).

Deep learning

Deep learning has emerged as a powerful technique for missing value imputation, particularly in the context of gene expression data and time series data. Deep learning has been applied to missing value imputations in various domains, including gene expression and time series data. Using deep neural networks and other deep learning models has shown promise in improving imputation accuracy and handling missing values effectively. These approaches leverage the power of deep learning algorithms to learn patterns and relationships from available data to impute missing values accurately (Park et al., 2022).

The optimal machine learning approach for imputing missing data varies depending on factors such as the data type, the extent of missing data, and the objectives of the analysis. As a general guideline:

- **K-Nearest Neighbors (KNN)** is suitable for straightforward datasets with minimal missing data.
- **Regression** is practical when dealing with datasets with a moderate amount of missing data.
- **Decision trees, random forests, and deep learning** excel in handling datasets with substantial missing data.

Recognizing that there is not a universal solution for missing data imputation is crucial. The choice of method should align with the specific dataset and analysis goals. It is often wise to explore multiple methods and determine which yields the best results for a given scenario.

Related works

In handling missing data, various financial research endeavors have explored an assortment of imputation methods drawn from both machine learning and statistical paradigms. These methods are pivotal in enhancing the integrity of financial data analyses, bolstering predictive models, and ensuring robust decision-making. This table presents a selection of articles that delve into imputation techniques for financial datasets. Each article investigates distinct methodologies, evaluates their effectiveness, and sheds light on their impact in addressing the ubiquitous challenge of missing values in financial data.

Table 3. Related works

Article Title	Year	Imputation Method(s)	Main Findings
Lan et al., "Multivariable data imputation for the analysis of incomplete credit data	2020	Bayesian algorithm	The experimental findings from this study indicate that BNII (Bayesian Network Imputation with Interaction Information) outperformed other widely recognized imputation methods by a significant margin. These results suggest that the proposed approach can potentially enhance the overall performance of a credit scoring system. Furthermore, this method could also be extended to enhance the effectiveness of other expert and intelligent systems.
Jadhave et al., "Comparing Imputation Methods for Financial Data Analysis"	2019	mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm. nob), and random sample	The analysis results demonstrate that the kNN (K-Nearest Neighbors) imputation method consistently outperforms other methods. Notably, the performance of this data imputation method remains consistent across different datasets and varying percentages of missing values within those datasets.
Stavseth et al., "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data."	2019	Multiple imputation techniques encompass expectation-maximization with bootstrapping, multiple correspondence analysis, latent class analysis, hot deck imputation, and multivariate imputation using chained equations with two distinct model specifications: logistic regression and random forests.	All methods exhibited relatively strong performance in scenarios with a substantial sample size (n=1000). However, when dealing with smaller sample sizes (n=200), the accuracy of regression estimates became highly dependent on the extent of missing data. Notably, when the proportion of missing data reached or exceeded 20%, several methods, including complete case analysis, hot deck, and random forests, produced biased estimates with insufficient coverage. In contrast, multiple imputation employing multiple correspondence analysis consistently demonstrated the most robust and reliable performance across different levels of missing data.
Ratolojanahary et al., Model selection to	2019	e Random Forest (R.F.), Boosted Regression Trees	The results obtained indicate that combining MICE (Multiple Imputation by Chained Equations) with machine learning techniques

improve multiple imputation for handling high rate missingness in a water quality dataset		(BRT), K- Nearest Neighbors (KNN) and Support Vector Regression (SVR).	such as SVR (Support Vector Regression), KNN (K-Nearest Neighbors), R.F. (Random Forest), and BRT (Boosted Regression Trees) yields superior performance compared to using MICE by itself. The MICE-SVR hybrid approach also strikes a favorable balance between performance and computational efficiency, making it an attractive choice.
Tang, and Ishwaran, " Random Forest Missing Data Algorithms "	2018	Random Forest Imputation,	The research findings suggest that R.F. (Random Forest) imputation is generally robust, and its performance improves as the degree of correlation in the data increases. R.F. imputation performed well even in scenarios with moderate to high levels of missing data and, in some cases, when the missing data was not missing at random (MNAR).
Schouten et al., "Generating missing values for simulation purposes: a multivariate amputation procedure."	2018	Nearest Neighbors (KNN), Decision Tree, and Bayesian Networks	The results indicate that Bayesian exhibits the most promising and favorable performance among the three classifiers tested.
Tutz and Ramzan "Improved methods for imputing missing data by nearest neighbor methods."	2015	weighted nearest neighbor	The study demonstrates that this method produces more minor imputation errors than other nearest-neighbor estimation techniques.
Ting "A comparison of multiple imputations with E.M. algorithm and MCMC method for quality of life missing data."	2010	Expectation-Maximization (E.M.) algorithm and Monte Carlo Markov chain (MCMC)	The research findings indicate no statistically significant difference in performance between the E.M. (Expectation-Maximization) algorithm and the MCMC (Markov Chain Monte Carlo) method. Moreover, the accuracy rates remained relatively high as the proportions of missing data increased. While the number of items used for imputation did have some impact on imputation accuracy, its influence was less substantial than initially anticipated.

Research Methodology

This paper examined four different machine learning methods for imputing missing values in time series data of S&P 500 companies and Bitcoin. The dataset covers the period from January 4, 2016, to June 2, 2023, and includes 425 variables with a time series length of 1900. Among the variables, the Bitcoin variable had 45 missing values, which were randomly removed from the data to test the accuracy of the methods. Changing the number of deleted data could potentially influence the results. However, our intent was not to exhaustively explore the sensitivity of the methods to varying degrees of data deletion. Instead, we emphasized comparing the inherent capabilities of machine learning and statistical approaches in handling missing data scenarios. To impute the missing values, we employed eight different methods: Nearest Neighbor Imputation, Deep Learning Imputation, Decision Tree Imputation, Random Forest as machine learning methods, and Mean/Mode Imputation, Regression, Hot-Deck, Gaussian Mixture Model and Expectation Maximization as statistical imputation methods. These methods were implemented using popular libraries such as Pandas and TensorFlow in Python. We accessed the S&P 500 and Bitcoin historical data from the Yahoo Finance database using the yfinance Python library to initiate our investigation. The datasets spanned from 2016 to 2023, providing a substantial temporal scope for our analysis. The daily frequency of the data was chosen to capture the inherent volatility and dynamics of financial markets.

In research, when implementing a model, tuning hyperparameters to achieve optimal performance is often necessary. One popular approach for hyperparameter tuning is Random Search. Unlike exhaustive methods like grid search, which systematically explores all possible combinations of hyperparameters within predefined ranges, Random Search selects parameter combinations randomly. This random sampling allows it to cover a wide range of values efficiently without the computational burden of exhaustively searching through all combinations. As a result, Random Search is significantly faster, particularly for models with numerous hyperparameters or when computational resources are limited. Despite its randomness, Random Search yields good results because it explores diverse parameter combinations, often discovering high-performing configurations that more deterministic approaches might miss. Thus, Random Search is a valuable tool in the researcher's arsenal for efficiently optimizing model performance.

Figure 2 illustrates the fundamental concept of the analysis. Initially, we start with the original datasets containing all the values. Subsequently, we

introduce missing values into the data following a Missing Completely at Random (MCAR) assumption, where 45 data points out of 1900 in the Bitcoin variable are intentionally deleted. These missing values are then replaced using the eight methods mentioned. Finally, we evaluate the difference between the replaced and original values using the Mean Absolute Percentage Error (MAPE).

MAPE, an acronym denoting Mean Absolute Percentage Error, is one of the most commonly employed methods for calculating forecasting accuracy. It offers a convenient and effective way to assess accuracy as it provides a readily interpretable measure. The formula to calculate MAPE is as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i}$$

A_i is the actual value, F_i is the forecast value, and n is the number of samples.

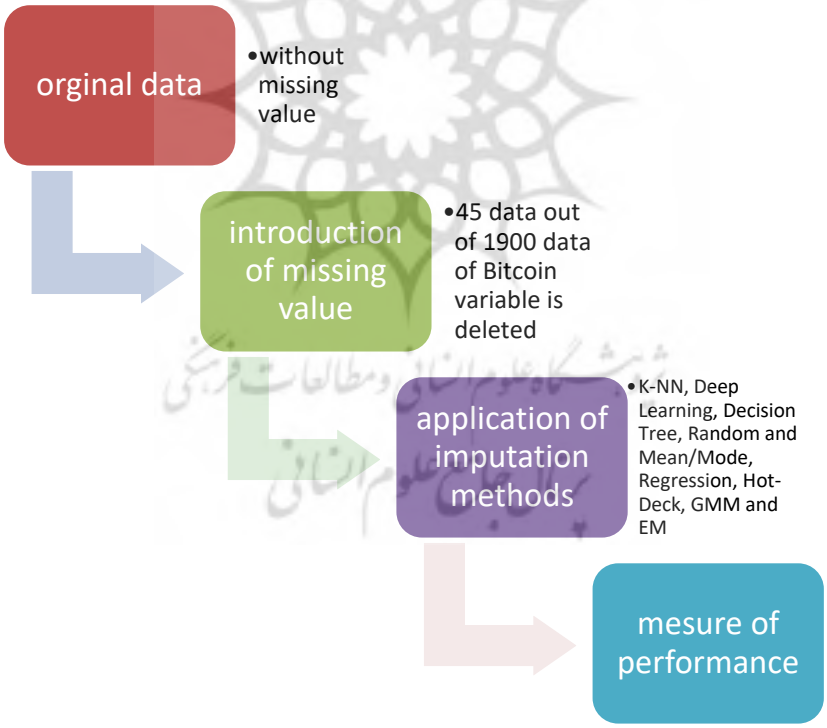


Figure 2. The general principle of the analysis

Results

In this research, we used eight popular statistical methods and machine learning to impute missing data to calculate the missing values. As mentioned, these methods are chosen because of their popularity among articles. The following diagram shows the result of a simple search in Google Scholar for the first eight months of 2023 with the keyword imputation of missing values and imputation methods.

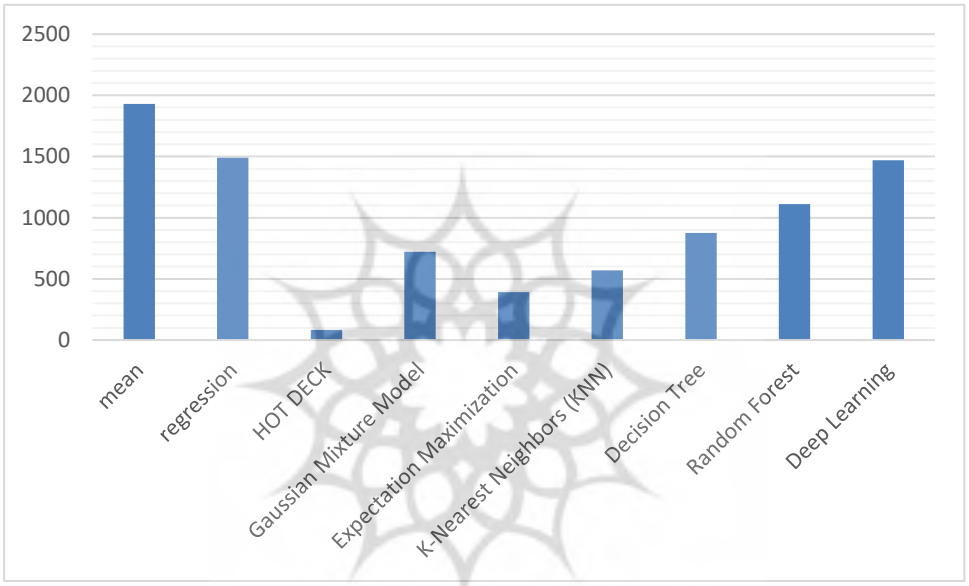


Chart 1. The result of google scholar search

Source: Google Scholar

Chart 2 and 3 plot the performances of each method as statistical and machine learning methods to impute missing values. As expected, machine learning methods have the power to predict missing values. Among the Machin learning methods, Random Forest performs best to impute the missing values. Chart 2 depicts a close alignment between the original data lines and the values forecast by the random forest model.

The best imputer of statistical methods is hot-deck. Hot-deck imputation is a method for filling in missing data by relying on similar cases within a dataset. It identifies comparable cases with available information and uses their data to estimate and replace the missing values.

Chart 2. Machin learning methods

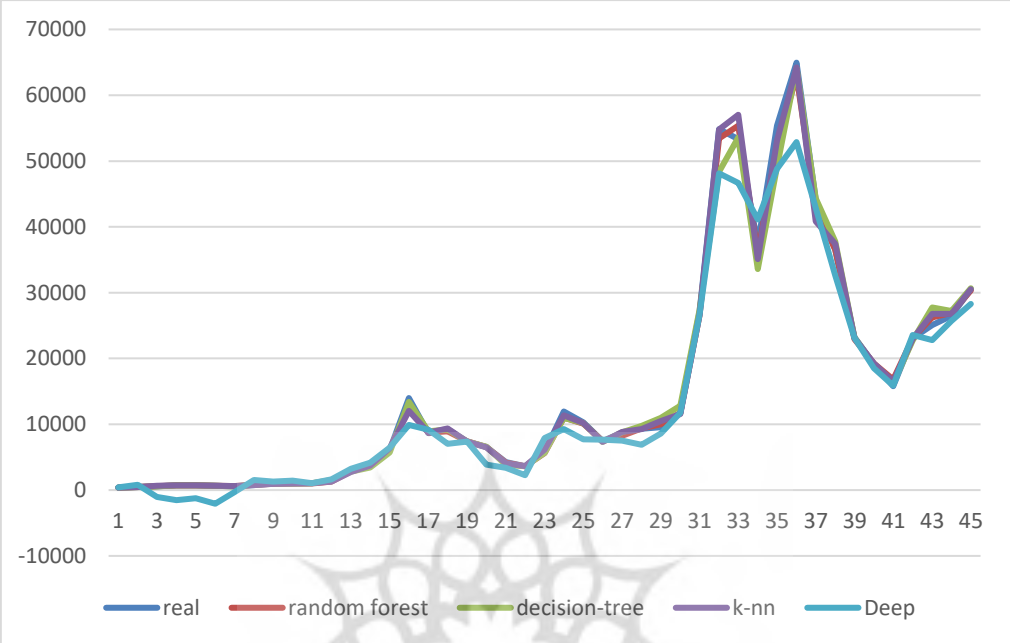
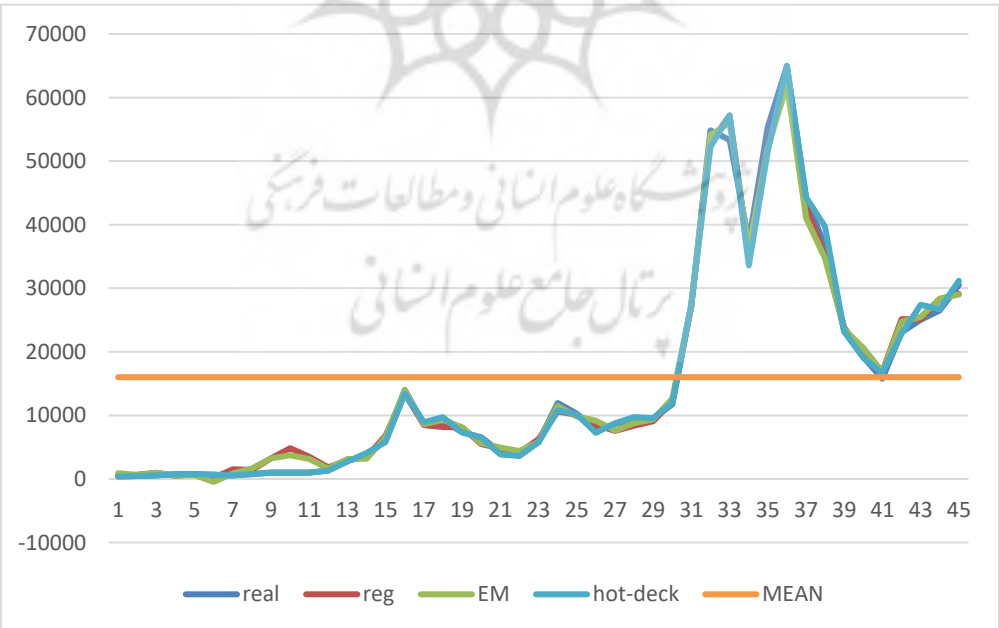


Chart 3. Statistical methods



Each imputation method's performance was assessed based on prediction accuracy. After completing the missing value imputation process, the final step involves evaluating the imputation results. A commonly used approach is directly comparing the collected dataset's original values with the estimated or predicted values. One such evaluation metric is MAPE, which stands for Mean Absolute Percentage Error.

MAPE serves as a measure of the accuracy of a forecasting method. It calculates the average of the absolute percentage errors for each entry in a dataset, providing insight into how accurately the forecasted quantities align with the actual quantities. MAPE is particularly effective for analyzing large datasets and necessitates the presence of non-zero values in the dataset. Lower MAPE values indicate a model's better ability to predict values.

In this study, we utilized Mean Absolute Percentage Error (MAPE) as the metric to gauge the accuracy of the imputation methods employed for handling missing values. The results obtained are presented in table 4:

Table 4. Accuracy of imputation methods for missing values

Method	MAPE (%)
Nearest Neighbor Imputation	3.395468
Deep Learning Imputation	47.2042
Decision Tree Imputation	4.271175
Random Forest Imputation	3.001032
mean	670.1858
hot-deck	3.455257
GMM	87.49537
EM	34.82417

As observed, the MAPE values for all four methods are relatively close. Among the eight imputation methods, the Random Forest Imputation demonstrates the highest level of accuracy. The mean imputation is the worst method to predict missing values.

Upon careful analysis, Random Forest Imputation emerged as the standout performer, demonstrating the highest level of accuracy among the evaluated methods. This finding underscores the efficacy of Random Forest Imputation approaches in handling missing data. Random forests excel in imputing missing values due to their ability to capture complex relationships within the data, leveraging an ensemble of decision trees to make robust predictions. By harnessing the collective intelligence of multiple decision trees, Random Forest Imputation effectively mitigates the impact of missing data on predictive accuracy. Conversely, mean imputation, despite its simplicity, emerged as the

least effective method. Mean imputation suffers from inherent limitations, primarily stemming from its reliance on a single summary statistic to estimate missing values. This approach must be revised to capture the nuanced patterns and variability in the data, leading to inaccuracies in imputed values. Moreover, mean imputation disregards the underlying structure of the data, potentially introducing bias and distorting the proper distribution of the variables.

The stark contrast between Random Forest Imputation and mean imputation underscores the importance of employing sophisticated techniques capable of capturing the inherent complexity of real-world datasets. While mean imputation may offer a quick fix, its shortcomings become evident in scenarios where data patterns are more intricate. In contrast, Random Forest Imputation excels in discerning underlying patterns and making informed predictions, thereby enhancing the reliability and robustness of imputed data.

Conclusion

This study investigated the effectiveness of imputing missing values using a combination of machine learning and statistical methods within financial data analysis. Leveraging a real-world dataset encompassing daily data for S&P 500 companies and Bitcoin spanning from 2016 to 2023, our research sought to empirically evaluate the performance of various imputation techniques across diverse financial datasets. Through this exploration, we aimed not only to deepen our understanding of their applicability but also to shed light on the strengths and limitations of these strategies, thus advancing our knowledge in this critical domain. Upon rigorous analysis, it becomes apparent that the random forest method surpasses all counterparts, exhibiting the most effective imputation results. This finding is consistent with prior studies by Ratolojanahary et al. (2019) and Tang and Ishwaran (2018). However, it is noteworthy to acknowledge the cautionary findings of Stavseth et al. (2019), who revealed potential biases and inadequate coverage in estimates when the proportion of missing data exceeded certain thresholds, particularly in the case of complete case analysis and random forests.

Conversely, the mean imputation method demonstrates the least satisfactory outcomes, aligning with the observations made by Zhang (2016). While mean imputation offers simplicity and convenience, its disregard for data relationships, distribution, and variability can lead to biased and inaccurate imputed values, thereby compromising the integrity of subsequent analyses and model performance. This assessment underscores the pivotal role of machine

learning techniques, notably the random forest algorithm, in addressing challenges associated with missing data while indicating a comparative lag in the performance of traditional statistical methods. The results of this study resonate with the findings of Jariz et al. (2010), further emphasizing the significance of employing advanced machine-learning approaches for accurate data imputation and advocating for their strategic integration across diverse analytical contexts.

This study bridges the gap between machine learning and statistical techniques for missing value imputation in financial contexts. The demonstrated excellence of machine learning approaches opens avenues for conducting more resilient and precise analyses within the field, thereby contributing to refining financial strategies and advancing predictive models. While our study provides valuable insights into the efficacy of imputation methods, further research is warranted to explore the robustness of these techniques across varying datasets and analytical scenarios. Additionally, investigating potential thresholds for acceptable levels of missing data and their implications on imputation accuracy would enhance the applicability of findings in practical settings. Moreover, conducting comparative studies on different machine learning algorithms for imputation could offer deeper insights into their relative performance and suitability for specific contexts. Finally, longitudinal studies tracking the performance of imputation methods over time contribute to the ongoing refinement and optimization of data imputation strategies in financial analyses.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest concerning the research, authorship and, or publication of this article.

Funding

The authors received no financial support for the research, authorship and, or publication of this article.

References

Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. International Journal of Advanced Computer Science and Applications, 9(6).

Acuna E, Rodriguez C (2004). The treatment of missing values and its effect on the classifier accuracy. In: Banks D et al. (eds) Classification, clustering, and data mining applications. Springer, Berlin, pp 639–648

Cham, H., & West, S. G. (2016). Propensity score analysis with missing data. *Psychological methods*, *21*(3), 427.

Chen, Y. C. (2022). Pattern graphs: a graphical approach to nonmonotone missing data. *The Annals of Statistics*, *50*(1), 129–146.

Dantan, E., Proust-Lima, C., Letenneur, L., & Jacqmin-Gadda, H. (2008). Pattern mixture models and latent class models for the analysis of multivariate longitudinal data with informative dropouts. *The International Journal of Biostatistics*, *4*(1)

Dany'el Irawan, N., Wijono, W., & Setyawati, O. (2017). Perbaikan missing value menggunakan pendekatan korelasi pada metode k-nearest neighbor. *Jurnal Infotel*, *9*(3), 305-311.

Demirtas, H. (2018). Flexible imputation of missing data. *Journal of Statistical Software*, *pp. 85*, 1–5.

Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*, 1-17.

Husson, F., Josse, J., Narasimhan, B., & Robin, G. (2019). Imputation of mixed data with multilevel singular value decomposition. *Journal of Computational and Graphical Statistics*, *28*(3), 552-566.

Ismail, A.R., Abidin, N.Z., & Maen, M.K. (2022). Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare. *Journal of Robotics and Control (JRC)*.

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, *33*(10), 913–933.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., &

Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, *50*(2), 105-115.

Kang, H. (2013). The prevention and handling of the missing data. *Korean journal of anesthesiology*, *64*(5), 402–406.

Lakshminarayan, K., Harp, S. A., & Samad, T. (1999). Imputation of missing data in industrial databases. *Applied Intelligence*, *11*(3), 259-275.

Lan, Q., Xu, X., Ma, H., & Li, G. (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems with Applications*, *141*, 112926.

Lin, T. H. (2010). A comparison of multiple imputation with E.M. algorithm and MCMC method for quality of life missing data. *Quality & Quantity*, *pp. 44*, 277–287.

Lin, W. C., & Tsai, C. F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, *53*, 1487-1509.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological methods & research*, *18*(2-3), 292-326.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 22–36.

Moeinol, H. H., et al. (2022). An approach for handling missing values in classification tasks. International Journal of Engineering and Technology, 14(2), 133-142.

Moinul, M., Amin, S. A., Kumar, P., Patil, U. K., Gajbhiye, A., Jha, T., & Gayen, S. (2022). Exploring sodium glucose cotransporter (SGLT2) inhibitors with machine learning approach: A novel hope in anti-diabetes drug discovery. *Journal of Molecular Graphics and Modelling*, *111*, 108106

Park, J., Müller, J., Arora, B., Faybishenko, B., Pastorello, G., Varadharajan, C. & Agarwal, D. (2023). Long-term missing value imputation for time series data

using deep neural networks. *Neural Computing and Applications*, *35*(12), 9071-9091.

Rahman, M.G., & Islam, M.Z. (2011). A Decision Tree-based Missing Value Imputation Technique for Data Pre-processing. *Australasian Data Mining Conference.*

Rahman, M.G., & Islam, M.Z. (2014). iDMI: A novel technique for missing value imputation using a decision tree and expectation-maximization algorithm. *16th Int'l Conf. Computer and Information Technology*, 496-501.

Ratolojanahary, R., Ngouna, R. H., Medjaher, K., Junca-Bourié, J., Dauriac, F., & Sebilo, M. (2019). Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications*, *131*, 299-307.

Raymond M, Roberts D (1987). A comparison of methods for treating incomplete data in selection research. Educ Psychol Meas 47:13–26

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Saha, S., Ghosh, A., Bandopadhyay, S., & Dey, K.N. (2019). Missing value imputation in DNA microarray gene expression data: a comparative study of an improved collaborative filtering method with decision tree based approach. *Int. J. Comput. Sci. Eng., pp. 18*, 130–139.

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909-2930.

Sidek, R. M., et al. (2016). A review of missing value imputation methods for time series. Annual Research & Review in Biology, 10(6), 1-9.

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE open medicine*, *7*, 2050312118822912.

Strike K, Emam KE, Madhavji N (2001). Software cost estimation with incomplete data. IEEE Trans Softw Eng 27(10):890–908

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *10*(6), 363–377.

Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. *Statistical*

Analysis and Data Mining: The ASA Data Science Journal, *10*(6), 363–377.

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, *90*, 84-99.

Wang, Z., Zhao, B., Guo, H., Tang, L., & Peng, Y. (2019). Deep Ensemble Learning Model for Short-Term Load Forecasting within Active Learning Framework. Energies, 12(20), n/a.

Zhang, Y., Li, M., Wang, S., Dai, S., Luo, L., Zhu, E. & Zhou, H. (2021). Gaussian mixture model clustering with incomplete data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(1s), 1–14.

Zhang, Y., Liu, J., Liu, H., Lu, Y., Wang, S., & Zhai, Y. (2022). High Dimensional Missing Data Imputation for Classification Problems: A Hybrid Model based on K-Nearest Neighbor and Genetic Algorithm. *2022 International Symposium on Advances in Informatics, Electronics and Education (ISAIEE)*, pp. 572–578.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).

Bibliographic information of this paper for citing:

Goldani, Mahdi (2024). Comparative Analysis of Missing Values Imputation Methods: A Case Study in Financial Series (S&P500 and Bitcoin Value Data Sets). *Iranian Journal of Finance*, 8(1), 47-70.