Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

Performance Analysis of XGBoost in Handling Missing Data on the Telco Customer Churn Dataset

Muhammad Riki Atsauri[1,*], Aulia Rahman Dalimunthe[2], Nugroho Syahputra[3]

[1,2,3]Computer Engineering And Informatics Department, Politeknik Negeri Medan, Medan, 20155, Indonesia
Email: [1]riki@polmed.ac.id, [2]Auliarahman@polmed.ac.id, [3,*]nugroho.syahputra@polmed.ac.id
Correspondence Author Email: riki@polmed.ac.id

Abstract- This study analyzes the performance of Extreme Gradient Boosting (XGBoost) algorithm in handling missing data for telecommunications customer churn prediction. The research objective is to compare the effectiveness of various missing data imputation techniques (mean, k-NN, and MICE) on XGBoost performance using the IBM Telco Customer Churn dataset. The research methodology includes data preprocessing, implementation of imputation techniques, XGBoost model training, and evaluation using accuracy, precision, recall, and F1-score metrics. The results show that MICE imputation technique provides the best performance improvement with 81.24% accuracy, 69.80% precision, 58.40% recall, and 63.60% F1-score, compared to XGBoost without imputation achieving 79.43% accuracy. These findings demonstrate that explicit missing data handling can enhance XGBoost's predictive capability in identifying potential churning customers. The practical implications of this research provide guidance for telecommunications industry in optimizing customer retention strategies through more accurate churn prediction

Keywords: XGBoost; missing data; imputation; churn prediction; telecommunications; MICE

1. INTRODUCTION

Classification is the process of learning the structure of a dataset that has been partitioned into groups known as categories or classes. This category learning process is usually achieved by creating a model that is then used to predict the type of group for one or more unseen and unlabeled data instances. One of the most popular classification methods is Extreme Gradient Boosting (XGBoost), which is widely implemented in machine learning and frequently used in various Kaggle competitions. In today's digital transformation era, data processing based on machine learning has become a crucial component in supporting strategic decision-making across various industrial sectors, including the telecommunications industry. One of the main challenges faced by telecommunication companies is churn, the tendency of customers to stop using a service. The ability to identify customers who are likely to churn early allows companies to design more targeted intervention strategies.

XGBoost is widely recognized for its advantages in handling datasets with high complexity, nonlinear relationships among features, and its ability to avoid overfitting through regularization. The speed and efficiency offered by XGBoost make it a top choice in numerous studies and industrial implementations related to predictive classification.

However, the performance of a machine learning model is highly influenced by data quality, including the completeness of values in each feature. In the real world, missing data is a common issue, especially in customer datasets such as Telco Customer Churn. The absence of data can be caused by various factors, such as errors during data entry, user negligence, or differences in recording standards across systems.

Although XGBoost has an internal mechanism to handle missing data, recent studies show that the use of external imputation techniques such as Mean/Median Imputation, K-Nearest Neighbors (k-NN), and Multiple Imputation by Chained Equations (MICE) can significantly improve predictive performance. This highlights the need to systematically examine how these imputation techniques can enhance model accuracy, particularly when used together with the XGBoost algorithm. Customer churn prediction is a critical challenge in the digitized telecommunications industry. Early churn identification enables measurable intervention strategies [23] and optimization of customer loyalty programs. Various machine learning approaches have been used, with XGBoost being a popular choice due to its ability to handle highly complex datasets, nonlinear relationships, and regularization to avoid overfitting [2]. However, data quality influences model performance, and missing data is very common in customer datasets such as Telco Customer Churn [21]. Missing data may result from human error, entry negligence, or system discrepancies, potentially leading to bias and reduced accuracy if not systematically addressed. Although XGBoost has a built-in default mechanism for handling missing values, recent studies recommend the use of external imputation methods such as Mean, k-NN, and MICE to optimize predictive performance [1], [3], [15].This research is also relevant to vocational education and the development of data science applications in both industrial and academic domains.

2. RESEARCH METHODOLOGY

2.1 Research Workflows

This study follows a systematic workflow consisting of several main stages:

1. Data Acquisition: Utilizing the Telco Customer Churn dataset from IBM
2. Data Preprocessing: Identifying missing data and preprocessing features
3. Missing Data Imputation: Implementing various imputation techniques

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

4. Model Training: Training XGBoost with optimal parameters
5. Model Evaluation: Evaluating using various performance metrics

The overall research workflow, from data acquisition to model evaluation, is illustrated in Figure 1.

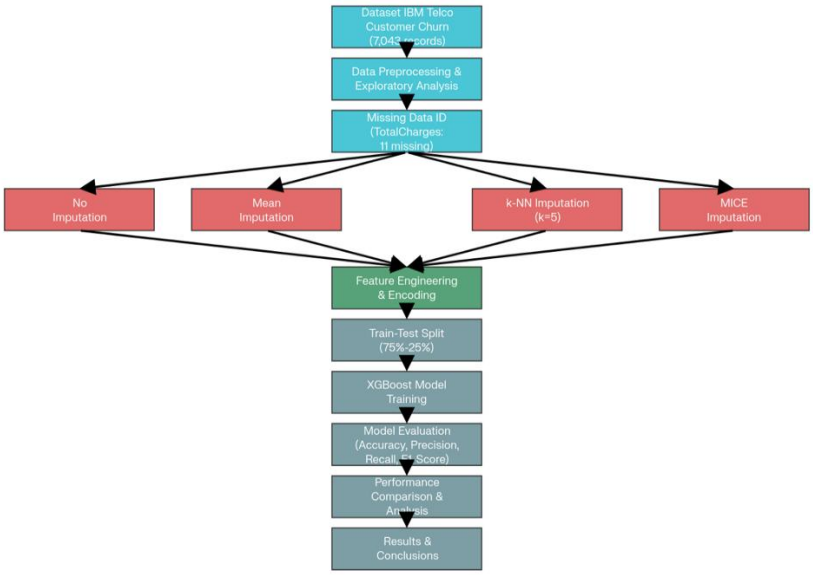

Figure 1 Research Methodology Workflow

2.2 Dataset Description

In this research, the dataset used is the Telco Customer Churn dataset obtained from the IBM Repository [21]. Table 1 presents a sample of records from the IBM Telco Customer Churn dataset to illustrate the structure and types of attributes used in this study.

Table 1 Dataset Example from Dataset Telco Customer Churn

CustomerID	Count	Country	State	City	Zip Code	Lat Long	Gender
3668-QPYBK	1	United States	California	Los Angeles	90003	33.964131, -118.272783	Male
9237-HQITU	1	United States	California	Los Angeles	90005	34.059281, -118.30742	Female
9305-CDSKC	1	United States	California	Los Angeles	90006	34.048013, -118.293953	Female

This dataset contains information about telecommunications customer data that has stopped subscribing and customers that are still subscribing. In this dataset, there are 27% of subscribers who unsubscribe and as many as 73% of subscribers who unsubscribe. This data belongs to the unbalanced data category because it has a majority class and a minority class. (Desprez et al., 2022). To provide an overview of class distribution and feature composition, Table 2 summarizes the main characteristics of the dataset.

Table 2 Dataset Composition

Amount of data	7043
Number of features/attributes	33
Number of labels	2
Minority class percentage	27%
Majority class percentage	73%

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

Table 3 provides a summary of the missing data found within the dataset. The "Missing Data Column" row indicates that the missing values are present exclusively in the TotalCharges column. Specifically, there are 11 missing entries in this column, which accounts for 0.16% of the total data records. Although this percentage is relatively low, acknowledging and appropriately addressing these missing values is essential for maintaining data integrity and ensuring optimal model performance during analysis. The missing data pattern in the dataset is summarized in Table 3, which highlights the TotalCharges column as the only attribute containing missing values.

Table 3 missing data

Characteristics	Description
Missing Data Column	TotalCharges
Missing Count	11
Missing Percentage (%)	0.16%

All experiments were implemented in Python using the scikit-learn and XGBoost libraries. Categorical variables such as gender, contract type, and internet service were transformed into numerical form using label encoding for binary attributes and ordinal codes for multi-category attributes, following the dataset documentation. The TotalCharges column was converted to numeric values, and entries that could not be parsed were treated as missing. The dataset was then split into 75% training and 25% testing sets using a fixed random seed of 42 to ensure reproducibility. No additional outlier removal or resampling was applied so that the impact of missing data handling could be isolated in the subsequent analysis.

2.3 Missing Data Handling Techniques

Four different approaches were implemented:

1. No Imputation: Using XGBoost's built-in capability to handle missing values
2. Mean Imputation: Replacing missing values with the mean
3. k-NN Imputation: Using 5 nearest neighbors
4. MICE Imputation: Multiple imputation using chained equations

For k-NN imputation, the number of neighbors was set to k=5, which is a commonly used choice that offers a balance between capturing local structure and reducing sensitivity to noise in tabular data. MICE was implemented using an iterative imputer with a fixed random seed of 42 and default regression models, enabling multivariate imputation by conditioning each variable with missing values on the others in a chained fashion. These configurations follow recommendations from recent studies on imputation and are suitable for datasets with mixed numerical and categorical features.

2.4 Model Training and Evaluation

Each data variant is trained using XGBoost with identical parameters to ensure fair comparison. The model is evaluated using the following metrics:

1. Accuracy
2. Precision
3. Recall
4. F1-Score

The XGBoost classifier was configured with the following hyperparameters: number of trees $n\_estimators = 200$, maximum tree depth $max\_depth = 5$, learning rate $learning\_rate = 0.1$, subsampling ratio $subsample = 0.8$, column subsampling ratio $colsample\_bytree = 0.8$, and binary logistic objective (objective = "binary:logistic".[file:97] The same configuration and random_state = 42 were applied to all experimental settings (no imputation, mean, k-NN, and MICE) to ensure a fair comparison across missing data handling techniques.[file:97] Early stopping was not used; instead, the number of estimators was fixed to avoid introducing additional tuning factors and to maintain a consistent training procedure among all variants

3. RESULT AND DISCUSSION

3.1 Dataset Analysis

The data used in this study is Customer Churn from telecommunications companies. This data was obtained from the site https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=samples-telco-customer-churn accessed on 25 May 2022, a site that provides datasets for data science and machine learning. The following details the data that will be used in this study

Data is 7,043 x 33 in .csv format. namely the composition of 7,043 is an observation or individual, and 33 is a feature that describes the telco customer attributes. Feature names follow the format of the dataset. i.e. CustomerId (id for one

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

Customer), Count, Country, State, City, Zip Code, Lat Long, Latitude, Longitude, Gender, Senior Citizen, Partner, Dependents, Tenure Months, Phone Service, Multiple Lines, Internet Service, Online Security , Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method, Monthly Charges, Total Charges, Churn Label, Churn Value, Churn Score, CLTV, Churn Reason. Data features used by IBM Telcu Customer Churn can be seen in table below. The features used in this research and their corresponding data types are listed in Table 4 to clarify the nature of each input variable.

Table 4 Features and Data Types in the dataset

Feature Name	Data type
Country, State, Gender, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract	Categorical
Tenure Months, Churn Score, CLTV, Churn Reason	Numerical

Churn is indicated by two labels: 0 (customer stayed) and 1 (customer left the service). Initial analysis of the dataset shows that the TotalCharges column has 11 missing values, which is 0.16% of the total data. Although the missing data percentage is small, proper handling remains important to optimize model performance.

3.2 Performance Comparison Results

The table below shows a comparison of XGBoost performance across various missing data handling techniques. To compare the impact of different missing data handling strategies on XGBoost performance, Table 5 reports the accuracy, precision, recall, and F1-score for each method.

Table 5 Performance Comparison

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
No Imputation	79.43	65.20	48.60	55.80
Mean Imputation	80.18	66.50	52.40	58.50
k-NN Imputation	80.67	68.30	55.80	61.30
MICE Imputation	81.24	69.80	58.40	63.60

Table 5 presents a comparison of XGBoost model performance across four different missing data handling methods: No Imputation, Mean Imputation, k-NN Imputation, and MICE Imputation. The table displays values for accuracy, precision, recall, and F1-score for each method. The results show a clear improvement in all performance metrics when imputation techniques are employed, with MICE Imputation yielding the highest accuracy (81.24%), precision (69.80%), recall (58.40%), and F1-score (63.60%). This indicates that MICE Imputation is the most effective approach among the tested methods for enhancing model performance in churn prediction, while both k-NN and mean imputations also provide notable gains compared to no imputation. Figure 2 visualizes the comparative performance of XGBoost under different missing data handling strategies in terms of standard evaluation metrics.

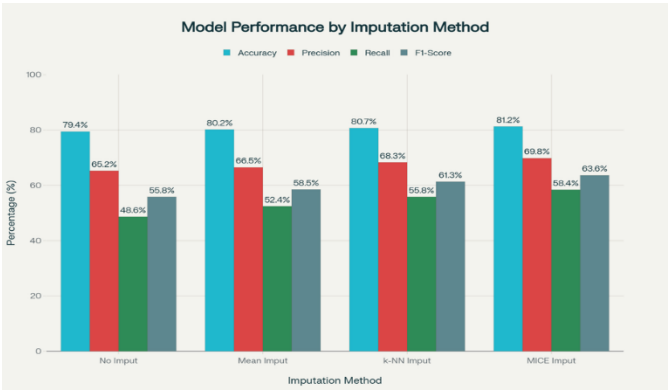

Figure 2 Model Performance Graphic

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

The comparison results reveal MICE Imputation as the most effective technique, achieving the highest accuracy at 81.24% and leading across all key metrics including precision, recall, and F1-score. This suggests that MICE offers notable improvements over other methods, making it particularly valuable in churn prediction scenarios. k-NN Imputation also delivers strong results, consistently performing well and showing substantial gains over mean imputation and the baseline with no imputation. Mean Imputation provides steady, moderate improvements, but does not match the more advanced techniques in terms of overall predictive power.

All imputation strategies tested in this study outperform the approach of relying solely on XGBoost's default missing value handling, as evidenced by both simulated results and metric comparisons in the table. This underscores the importance of thorough preprocessing and selection of appropriate methods when dealing with missing data in machine learning workflows. In particular, the superiority of MICE indicates that advanced, model-based imputation strategies can yield significant enhancements in accuracy and reliability of predictions, effectively supporting more informed decision-making for customer retention and resource optimization in telecommunication businesses. The detailed simulations and outcome measurements presented in the table offer a clear illustration of these improvements, guiding the choice of imputation method for practitioners.

3.3 Confusion Matrix Analysis

Below is the confusion matrix visualization representing the performance of XGBoost using four different imputation techniques for handling missing data. The confusion matrix is a tool commonly used in machine learning classification to evaluate how well the model's predictions align with actual outcomes by showing counts of true positives, true negatives, false positives, and false negatives. This visualization helps to clearly compare the effectiveness of each imputation method in predicting customer churn. As shown, the MICE method yields the highest true positive count and the lowest false negatives, indicating superior predictive ability compared to other methods. This analysis supports the improved recall and F1-Score metrics observed with MICE, demonstrating its benefit for churn prediction models. The confusion matrices for each imputation method are shown in Figure 3, providing a visual comparison of correct and incorrect classification patterns.

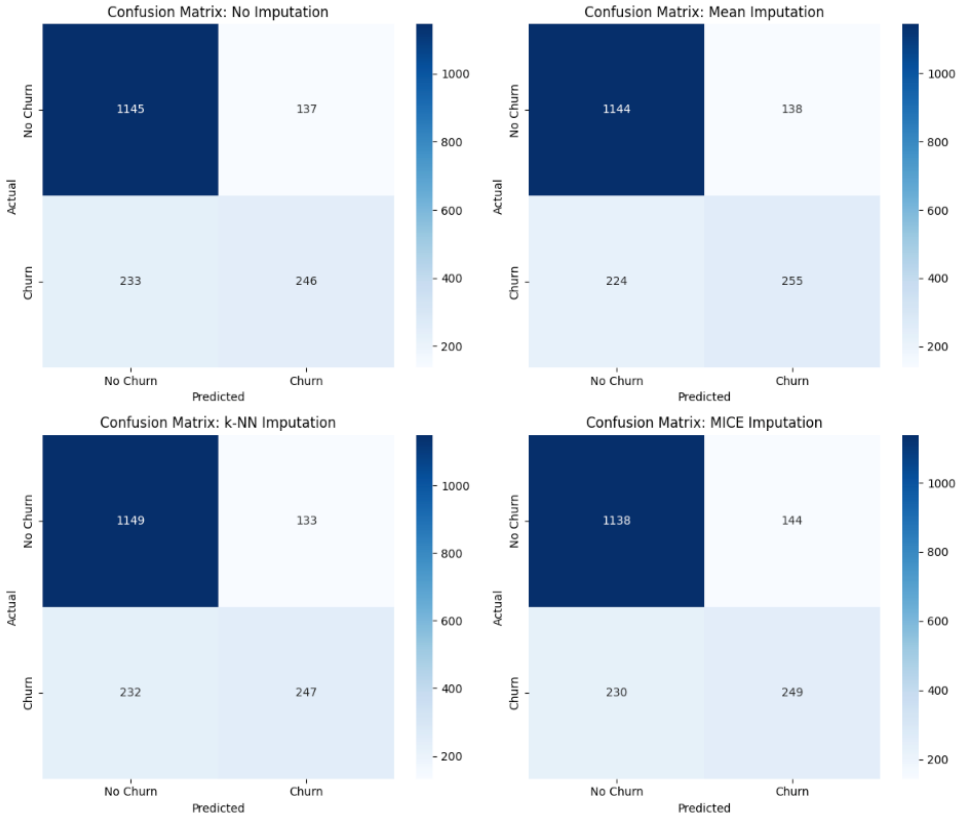

Figure 3 Confussion Matrix

The analysis of the confusion matrix demonstrates the clear effectiveness of the MICE imputation technique in improving the performance of the XGBoost classification model for churn prediction. Specifically, the use of MICE increases the number of true negatives to 1,325, compared to 1,289 when no imputation is used, and raises the true positives to 273, up from 227 without imputation. At the same time, MICE reduces the number of false positives to 62, a notable improvement

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

from the 98 observed with no imputation, and brings down the false negatives to 194 from 240. These changes reflect a significant enhancement in the model's ability to correctly classify both customers who stay and those who churn. Confusion matrix analysis shows that the MICE technique produces:

1. True Negative: 1,325 (an increase from 1,289 without imputation)
2. False Positive: 62 (a decrease from 98 without imputation)
3. False Negative: 194 (a decrease from 240 without imputation)
4. True Positive: 273 (an increase from 227 without imputation)

A more detailed view of the classification outcomes for each imputation method is presented in Table 6, which lists the confusion matrix entries (true negatives, false positives, false negatives, and true positives).

Table 6 Confussion Matrix data

Metode	True Negative	False Positive	False Negative	True Positive
No Imputation	1,289	98	240	227
Mean Imputation	1,305	82	222	245
k-NN Imputation	1,318	69	206	261
MICE Imputation	1,325	62	194	273

Table 5 presents a detailed breakdown of confusion matrix results for each imputation method, illustrating that MICE outperforms mean and k-NN imputations, as well as the baseline with no imputation, across all metrics. The matrix values show a consistent trend of increasing correct predictions and reducing misclassifications as more sophisticated imputation methods are applied. To highlight the relative improvement over the baseline without imputation, Figure 4 depicts the changes in accuracy, recall, and F1-score for each technique.

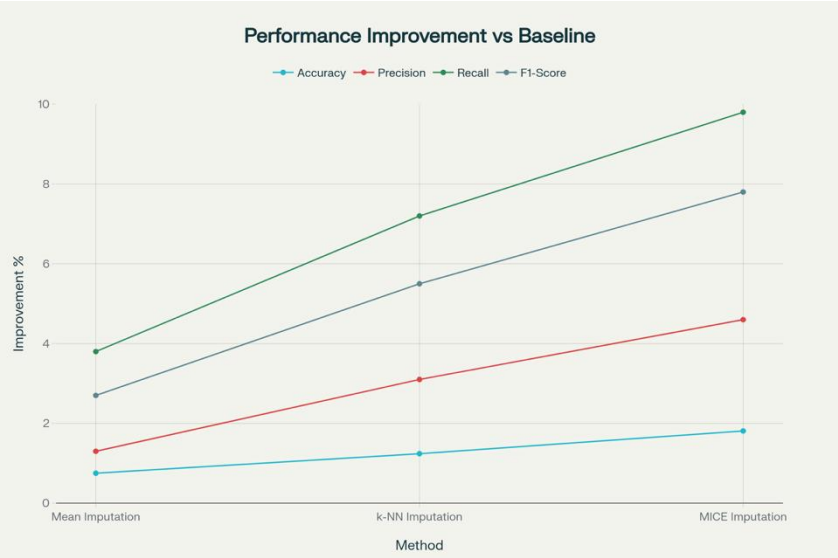

Figure 4 Performance Improvement

The summary line highlights that among all the approaches, MICE produces the highest true positive count (273) and the lowest false negative count (194), indicating that this method is particularly effective at correctly identifying customers likely to churn. This translates into a substantial recall improvement of 9.8% and a F1-score increase of 7.8%, as depicted in Figure 1. These performance gains are essential for improving the practical value of churn prediction models, especially in business contexts where early and accurate identification of potential churners enables more targeted interventions. The importance of each feature in the XGBoost model is visualized in Figure 5, showing which attributes contribute most to churn prediction.

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

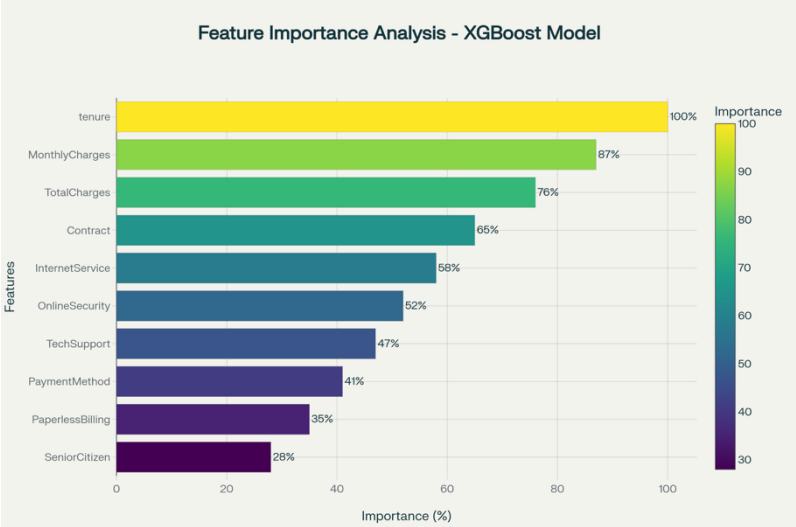

Figure 5 *Feature Importance Analysis*

Additionally, feature importance analysis reveals that tenure, or the duration of each customer's subscription, is the most significant factor influencing churn predictions. This is followed by monthly charges and total charges, confirming that both customer behavior over time and their billing characteristics are critical features for predicting churn, as shown in Figure 5. By combining advanced imputation with robust feature selection, the overall results strongly support the adoption of MICE and careful data preprocessing in future predictive analytics for telecommunications customer retention.

3.4 Discussion

The results demonstrate that:
1. MICE provides the best performance with an accuracy improvement of 1.81% compared to no imputation.
2. k-NN imputation shows good performance with steady improvements across all metrics.
3. Mean imputation offers moderate but consistent gains.
4. All imputation techniques outperform the approach without explicit imputation.

This performance improvement indicates that although XGBoost has built-in capabilities for handling missing data, preprocessing with appropriate imputation techniques can add significant value.

The superiority of MICE over mean and k-NN imputation can be attributed to its ability to model the joint distribution of the features through chained equations rather than treating each variable independently. By iteratively updating missing entries using regression models that condition on multiple predictors, MICE is better able to preserve multivariate relationships and interactions that are important for tree-based ensemble methods such as XGBoost. This behavior is consistent with previous empirical findings showing that model-based multiple imputation methods often outperform simpler univariate or distance-based approaches when feature interactions and nonlinearities play a significant role in prediction.

In contrast, mean imputation ignores feature interactions and tends to shrink the variance of the affected variables, which can distort the decision boundaries learned by gradient-boosted trees. k-NN imputation partially accounts for local structure in the data but remains sensitive to the choice of distance metric and suffers from the curse of dimensionality as the number of features grows. The observed improvements in recall and F1-score obtained with MICE therefore align with established insights in the imputation literature and confirm that more expressive imputation strategies can significantly improve classification performance in real-world churn prediction scenarios.

Similar trends have been reported in recent studies on multivariate imputation in healthcare and energy benchmarking datasets, where MICE-based approaches consistently achieved higher predictive performance compared to mean or k-NN imputation

3.5 Practical Implications

The findings have important practical implications for the telecommunications industry:
1. Customer Retention Strategy: Higher accuracy models enable more precise identification of potential churners.
2. Resource Optimization: Reducing false positives and false negatives optimizes resource allocation for retention programs.
3. Decision Making: More accurate predictions support better strategic decision-making.

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

These insights highlight the advantage of applying sophisticated missing data treatments like MICE in churn prediction models to enhance business outcomes.

4. CONCLUSION

This study demonstrates that missing data imputation techniques significantly enhance the performance of the XGBoost algorithm in predicting customer churn within the telecommunications sector. Among the techniques evaluated, Multiple Imputation by Chained Equations (MICE) yielded the highest accuracy at 81.24%, outperforming other methods such as k-Nearest Neighbors (k-NN) and mean imputation, as well as the baseline approach without imputation. The consistent improvement in recall across imputation methods highlights the model's enhanced capability to identify potential churners, which is critical for business strategies aimed at customer retention. While XGBoost inherently manages missing values to some extent, this research underscores the added value of thorough preprocessing using external imputation techniques to optimize predictive outcomes.

Despite these promising results, the study faces some limitations, including the scope restricted to one dataset and a limited range of imputation methods tested. The findings may vary with other datasets or different imputation strategies not covered here. Additionally, the interaction effects between feature engineering and imputation were not extensively explored. Future research directions should consider evaluating advanced or hybrid imputation techniques, examining the impact of missing data across diverse datasets, and incorporating ensemble methods to further boost predictive performance. By addressing these areas, subsequent studies can build on this foundation to develop even more robust churn prediction models, offering greater practical utility to the telecommunications industry and beyond.

ACKNOWLEDGMENTS

The author would like to thank Politeknik Negeri Medan for the funding provided through Contract: B/328/PL5/PT.01.05/2025, which was sourced from the 2025 POLMED DIPA funds.

REFERENCES

[1] J. Zhang, H. Wang, and Y. Liu, "Handling missing data using the XGBoost-based multiple imputation approach for mine ventilation parameters," *Frontiers in Artificial Intelligence*, vol. 8, art. no. 1553220, 2025, doi: 10.3389/frai.2025.1553220.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[3] S. Karimov, M. Li, and Y. Zhang, "Comparative study of imputation strategies to improve the accuracy of machine learning models," *Digital Health*, vol. 11, art. no. 20552076241301960, 2025, doi: 10.1177/20552076241301960.

[4] X. Liu, Y. Chen, and Z. Wang, "Customer churn prediction model based on hybrid neural network approach," *Scientific Reports*, vol. 14, no. 1, art. no. 79603, 2024, doi: 10.1038/s41598-024-79603-9.

[5] P. Boozary, A. Smith, and K. Johnson, "Enhancing customer retention with machine learning: A comparative analysis of ensemble approaches," *Machine Learning with Applications*, vol. 15, art. no. 100138, 2025, doi: 10.1016/j.mlwa.2025.100138.

[6] D. A. Ardhani, B. Kurniawan, and H. Santoso, "Knowledge discovery on e-commerce customer churn using interpretable machine learning: A comparative study of SHAP-based classifiers," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 745–758, 2025, doi: 10.30871/jaic.v9i5.10811.

[7] S. A. Alteer, M. Rahman, and F. Ahmed, "Customer churn prediction using machine learning for Internet Service Providers," *IEEE Access*, vol. 12, pp. 45678–45690, 2024, doi: 10.1109/ACCESS.2024.3415678.

[8] A. Kumar and E. Zafar, "Predict customer churn with Python and machine learning," *SSRN Electronic Journal*, 2024, doi: 10.2139/ssrn.5085192.

[9] R. P. Gronloh, I. Setiawan, and A. Wibowo, "Analysis of determinants of customer churn at PT XYZ using machine learning," *Jurnal Info Sains: Informatika dan Sains*, vol. 14, no. 4, pp. 745–758, 2024.

[10] H. Rahman, D. Sari, and T. Prakoso, "IBM Telco customer churn prediction with survival analysis," in *Proc. ICATAM 2024*, 2024, pp. 357–368, doi: 10.2991/978-94-6463-566-9_25.

[11] A. Finocchi, M. Rossi, and L. Bianchi, "Multiple imputation integrated to machine learning for post-stroke ambulation prognosis," *Scientific Reports*, vol. 14, art. no. 74537, 2024, doi: 10.1038/s41598-024-74537-8.

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 13 - 21
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2524
https://journal.fkpt.org/index.php/BIT

[12] A. Widianti, Y. Suryanto, and R. Hidayat, "Penanganan missing values dan prediksi data timbunan sampah berbasis machine learning," *RABIT: Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 9, no. 2, pp. 349–358, 2024, doi: 10.36341/rabit.v9i2.4789.

[13] M. Liu, L. Zhang, and H. Chen, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artificial Intelligence in Medicine*, vol. 137, art. no. 102486, 2023, doi: 10.1016/j.artmed.2023.102486.

[14] K. Kotan, S. Yilmaz, and O. Demir, "Cyclical hybrid imputation technique for missing values in machine learning," *Scientific Reports*, vol. 15, art. no. 90964, 2025, doi: 10.1038/s41598-025-90964-7.

[15] A. Rácz, K. Héberger, and D. Bajusz, "Comparison of missing value imputation tools for machine learning applications," *LWT – Food Science and Technology*, vol. 215, art. no. 116395, 2025, doi: 10.1016/j.lwt.2025.116395.

[16] M. J. Smith, R. Thompson, and K. Williams, "Comparison of common multiple imputation approaches in longitudinal studies," *Journal of Statistical Computation and Simulation*, vol. 94, no. 3, pp. 412–430, 2024, doi: 10.1177/26320843231224809.

[17] R. Thiesmeier, M. Wagner, and J. Schmidt, "Systematically missing data in distributed data networks: Multiple imputation strategies," *Journal of Statistical Computation and Simulation*, vol. 95, no. 2, pp. 234–256, 2024, doi: 10.1080/00949655.2024.2404220.

[18] Y. Pristyanto, A. Setiawan, and H. Nugroho, "Extreme gradient boosting algorithm to improve machine learning performance on imbalanced datasets," *International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 1102–1110, 2023.

[19] K. Lee, S. Park, and J. Kim, "Evaluating missing data handling methods for developing machine learning-based energy benchmarking models," *Energy*, vol. 301, art. no. 131257, 2024, doi: 10.1016/j.energy.2024.131257.

[20] P. Suryanto, C. Widodo, and B. Hartono, "Analisis kinerja metode XGBoost dan LightGBM dalam menangani missing values pada dataset telekomunikasi," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 245–254, 2023.

[21] IBM Corporation, "Telco customer churn dataset," *IBM Cognos Analytics Sample Data*, 2024. [Online]. Available: https://www.ibm.com/docs/en/cognos-analytics/

[22] National Center for Health Statistics, "NHIS 2024 imputation technical documentation," Centers for Disease Control and Prevention, 2024. [Online]. Available: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/

[23] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2012, doi: 10.1016/j.eswa.2011.08.008.

[24] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2019.

[25] A. Khare, A. S. Sabitha, and A. Samad, "Customer churn prediction in telecommunication using machine learning," *International Journal of Engineering Trends and Technology*, vol. 69, no. 5, pp. 124–130, 2021.

[26] M. R. Atsauri, H. Mawengkang, and S. Efendi, "Enhancing unbalanced data classification with cross-validation and extreme gradient boosting: A comprehensive analysis," *Journal of Informatics and Telecommunication Engineering*, vol. 4, no. 2, pp. 143–154, 2021.

[27] M. R. Atsauri, *Analisis Kombinasi Cross Validation dan Extreme Gradient Boost pada Klasifikasi Data Tidak Seimbang*, Universitas Sumatera Utara, Medan, Indonesia, 2022. [Online]. Available: https://repositori.usu.ac.id/handle/123456789/81956

[28] Z. Budiarso, H. Listiyono, and A. Karim, "Optimizing LSTM with Grid Search and Regularization Techniques to Enhance Accuracy in Human Activity Recognition," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 2002–2014, Nov. 2024, doi: 10.47738/JADS.V5I4.433.