Evaluating Imputation under Missingness Mechanisms for Credit Risk Modeling

Marcelo José Wecchi[1]

[1]UNIFESP - Universidade Federal de São Paulo
São José dos Campos – São Paulo – Brazil

marcelo.wecchi@unifesp.br

Abstract. *Missing data is a persistent challenge in credit risk modeling, where the mechanism generating missingness shapes both predictive performance and fairness. This study establishes a baseline by comparing five imputation methods - Mean, Median, Mode, KNN, and MICE - across three mechanisms (MCAR, MAR, MNAR) using the German Credit Dataset. Results show that MICE performs best under MCAR and MAR, while Mode proves more robust under MNAR. Fairness outcomes vary by mechanism, with MAR generally yielding lower disparities. Overall, findings demonstrate that no single imputation method is universally optimal, and that mechanism-aware selection is essential to balance accuracy and equity in credit risk applications.*

Resumo. *A falta de dados é um desafio persistente na modelagem de risco de crédito, onde o mecanismo que gera a falta de dados molda tanto o desempenho preditivo quanto a justiça. Este estudo estabelece uma linha de base ao comparar cinco métodos de imputação – Média, Mediana, Moda, KNN e MICE – em três mecanismos (MCAR, MAR, MNAR) usando o conjunto de dados de crédito alemão. Os resultados mostram que o MICE tem melhor desempenho sob MCAR e MAR, enquanto a Moda se mostra mais robusta sob MNAR. Os resultados de justiça variam de acordo com o mecanismo, com o MAR geralmente produzindo disparidades mais baixas. No geral, os resultados demonstram que nenhum método de imputação único é universalmente ideal e que a seleção consciente dos mecanismos é essencial para equilibrar a precisão e a equidade nas aplicações de risco de crédito.*

1. Introduction

Credit risk modeling guides lending decisions, but missing data remains a challenge in financial datasets. How missing values are handled affects both predictive accuracy and fairness in credit access, making imputation strategy a methodological choice with social impact [Khan et al. 2025].

Three canonical missingness mechanisms are recognized [Rubin 1976]:

1. **MCAR** (Missing Completely At Random): probability of missingness independent of observed or unobserved attributes.
2. **MAR** (Missing At Random): missingness depends on observed variables but not on the missing values themselves.
3. **MNAR** (Missing Not At Random): missingness depends on the missing variable itself or on unobserved factors.

Mechanism identification matters because methods effective under one may introduce bias under another [Khan et al. 2025].

Despite this foundation, practice often adopts mechanism-independent approaches, driven by computational convenience and limited empirical guidance in credit risk contexts [Leite 2024].

This study addresses two fundamental questions: (i) how do common imputation methods - Mean, Median, Mode, KNN, and MICE - perform under different mechanisms when applied to credit risk prediction? (ii) what are the fairness implications of imputation choices, and do methods that optimize accuracy also reduce demographic disparities?

Our contributions are threefold. We compare five imputation methods across three mechanisms using only the German Credit Dataset, establishing benchmarks. We evaluate performance through predictive and fairness metrics, reflecting deployment concerns. Finally, we provide mechanism-aware guidance for imputation selection, showing that optimal choices depend on the missingness mechanism and that no universal best method exists.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes methodology; Section 4 presents results; Section 5 discusses implications; and Section 6 concludes with recommendations.

2. Related Work

Research acknowledges that imputation choices affect both predictive accuracy and fairness in credit risk modeling; studies have examined missing data mechanisms, compared imputation methods, and explored fairness-related implications, yet the integration of these dimensions remains limited [Putri et al. 2025].

2.1. Missingness Mechanisms in Credit Data

Studies show that imputation performance varies across MCAR, MAR, and MNAR scenarios [Khan et al. 2025]. Methods tailored to MNAR, such as NRMPI [Li et al. 2025] and reject-aware approaches [Crook and Banasik 2004], explicitly model dependence on unobserved values. Other work highlights the use of binary indicators to capture missingness patterns [Zhang et al. 2026].

2.2. Imputation Method Comparisons

Comparative evaluations reveal that method effectiveness depends on dataset characteristics and missingness type. Median and KNN often outperform mean imputation [Leite 2024], while MICE achieves stronger results under lower missingness rates [Sue et al. 2024]. Localized estimates via KNN have shown gains over global central tendency imputers [Nogueira 2025]. Broader analyses confirm that optimal selection is mechanism- and context-dependent [Khan et al. 2025].

2.3. Fairness in Imputation

Imputation strategies can alter fairness outcomes. Caton *et al.* ranked methods by fairness metrics, showing that accuracy does not guarantee equity [Caton et al. 2022]. Zhang and Long documented cases where imputation disadvantaged specific groups despite unbiased missingness patterns [Zhang and Long 2021]. Frameworks incorporating

fairness-performance trade-offs emphasize the need to evaluate both dimensions jointly [Khan et al. 2025].

2.4. Research Gap

Previous studies have advanced the understanding of mechanisms, performance, and fairness in isolation, yet an assessment encompassing all three aspects remains scarce. We found no study that comprehensively compared multiple imputation methods under MCAR, MAR, and MNAR while jointly assessing predictive and fairness metrics on a standard benchmark. This study addresses that gap by evaluating five methods (Mean, Median, Mode, KNN, MICE) on the German Credit Dataset [Hofmann 1994], providing mechanism-aware, fairness-conscious guidance for imputation in credit risk modeling.

3. Methodology

The baseline design employs stratified cross-validation with logistic regression, measuring both predictive performance and fairness. This section describes the dataset characteristics, the simulation of missing data, the imputation methods, the evaluation protocol, and the metrics.

3.1. Dataset

The German Credit Dataset [Hofmann 1994] contains 1,000 records by 20 features. As it has no missing values, controlled simulations of missingness mechanisms were applied with known ground truth. For the protected attributes, we choose the age.

3.2. Missingness Simulation

The mechanisms were simulated to reflect real-world scenarios. Target features included duration, credit amount, installment rate, residence time, and number of existing credits.

Missingness Mechanisms Setup

MCAR Randomly applied to 20% of the target feature values.

MAR The probability depends on age, ranging from 0% for the youngest individuals (19 years old) to 60% for the oldest (75 years old), affecting the characteristics of the target.

MNAR Applied to values above the 97th percentile in the target columns, making the probability of missingness directly related to the magnitude of the unobserved value itself.

The values used for each of the mechanisms were obtained empirically, with the aim of achieving an overall missing data rate of $\approx 5\%$.

3.3. Imputation Methods

Five methods were evaluated: Mean, Median, Mode, KNN ($k = 5$), and MICE (10 iterations). These represent increasing complexity, from simple central tendency imputers to iterative regression-based approaches [van Buuren and Groothuis-Oudshoorn 2011, Leite 2024].

3.4. Evaluation Protocol

Experiments used 5-fold stratified cross-validation. Missingness was simulated on training folds, imputers were fit using training data, and applied to test folds. Logistic Regression with balanced class weights served as the classifier, addressing class imbalance.

3.5. Evaluation Metrics

Performance was assessed with AUC, F1, and Brier score. Fairness was measured with Statistical Parity Difference and Equalized Odds Difference [Caton et al. 2022, Hardt et al. 2016]. Imputation quality was evaluated with RMSE and KL-divergence, capturing point-wise accuracy and distributional preservation [Nastoska et al. 2025, Kwon et al. 2025]. Other metrics were not considered in this experiment.

4. Experimental Results

This section presents comprehensive findings from evaluating 15 experimental configurations: 3 missingness mechanisms (MCAR, MAR, MNAR) combined with 5 imputation methods (Mean, Median, Mode, KNN, MICE). Each configuration was assessed using 5-fold stratified cross-validation on the German Credit Dataset, measuring predictive performance through AUC-ROC, F1 Score, and Brier Score, alongside fairness metrics Statistical Parity Difference and Equalized Odds Difference.

4.1. Performance Overview

Table 1 presents the experimental results. MICE achieved the highest AUC predictive performance under MCAR and MAR mechanisms, while Mode imputation unexpectedly outperformed more sophisticated methods under MNAR. Mean and Median imputation showed performance degradation under MNAR compared to MCAR and MAR conditions.

4.2. Mechanism-Specific Performance

Figure 1 displays performance metric heatmaps across all configurations, revealing distinct patterns for each missingness mechanism. Under MCAR, imputation methods showed relatively uniform performance with MICE achieving marginal superiority (AUC=0.7309, F1=0.7394, Brier=0.2111). The MAR mechanism exhibited similar patterns with MICE maintaining the best performance (AUC=0.7298, F1=0.7409, Brier=0.2110). Under MNAR, Mode imputation unexpectedly achieved the highest F1 (0.7246) with competitive calibration (Brier=0.2150), while Mean and Median imputation suffered substantial degradation with AUC values dropping to approximately 0.71.

4.3. Distributional Preservation

Figure 2 presents kernel density estimates comparing the distributions of feature "credit_amount" with the complete original data. MICE and KNN imputation preserved distributional characteristics most faithfully under MCAR and MAR, maintaining multimodal structures and tail behaviors. Mean and Median imputation introduced artificial peaks at central tendency values, distorting the original distributions. Under MNAR, all imputation methods showed distributional distortions.

Table 1. Experimental results. Performance metrics: AUC-ROC and F1 Score (hight is better), and Brier Score (lower is better). Fairness metrics: SPD and EOdds, both measured as absolute values with lower indicating better fairness.

Mechanism	Imputer	AUC	F1	Brier	SPD	EOdds
MCAR	Mean	0.7299	0.7489	0.2117	0.2198	0.2159
MCAR	Median	0.7281	0.7488	0.2121	0.2252	0.2183
MCAR	Mode	0.7238	0.7402	0.2137	0.2199	0.2009
MCAR	KNN	0.7287	0.7461	0.2119	0.2202	0.2079
MCAR	MICE	0.7309	0.7394	0.2111	0.2205	0.2052
MAR	Mean	0.7265	0.7393	0.2119	0.1994	0.1941
MAR	Median	0.7260	0.7343	0.2123	0.1898	0.1867
MAR	Mode	0.7192	0.7274	0.2144	0.1967	0.1758
MAR	KNN	0.7286	0.7438	0.211	0.2181	0.2028
MAR	MICE	0.7298	0.7409	0.211	0.2001	0.1977
MNAR	Mean	0.7176	0.7156	0.2162	0.2092	0.1864
MNAR	Median	0.7160	0.7177	0.2166	0.2094	0.1862
MNAR	Mode	0.7193	0.7246	0.2150	0.2061	0.1882
MNAR	KNN	0.7220	0.7213	0.2138	0.2168	0.2041
MNAR	MICE	0.7230	0.7223	0.2138	0.2093	0.1934

4.4. Fairness Analysis

Figure 3 displays fairness metrics across all configurations using radar charts. Simple imputation techniques (mean, median and mode) achieved better demographic disparity and more equalized error rates between age groups for all missingness mechanisms. MAR configurations generally exhibited better fairness metrics than MCAR and MNAR, with Moda under MAR achieving EOdds = 0.1758, the lowest value across all configurations. However, this improved fairness came at the cost of reduced predictive performance, revealing a performance-fairness tradeoff.

4.5. Imputation Quality: RMSE and KL-Divergence

Beyond downstream predictive performance and fairness, we directly evaluate imputation quality through Root Mean Squared Error (RMSE) and Kullback-Leibler divergence (KL-divergence). RMSE measures point-wise reconstruction accuracy by comparing imputed values against ground truth on originally-missing positions, while KL-divergence quantifies distributional distortion between original and imputed feature distributions. These metrics provide complementary perspectives on imputation fidelity independent of downstream classification performance.

Table 2 presents mean RMSE values averaged across five numerical target features (duration, credit_amount, installment_commitment, residence_since, existing_credits) for all 15 experimental configurations. Under MCAR, MICE achieved the lowest RMSE (438.88), followed by KNN (483.69), demonstrating superior point-wise reconstruction accuracy. Mean (550.92), Median (572.68), and Mode (666.29) showed progressively higher errors, reflecting their inability to capture feature correlations. Under MAR, MICE achieved the lowest RMSE (459.59), marginally outperforming KNN (566.00), while simple imputers again showed substantially higher errors. Under MNAR, all methods

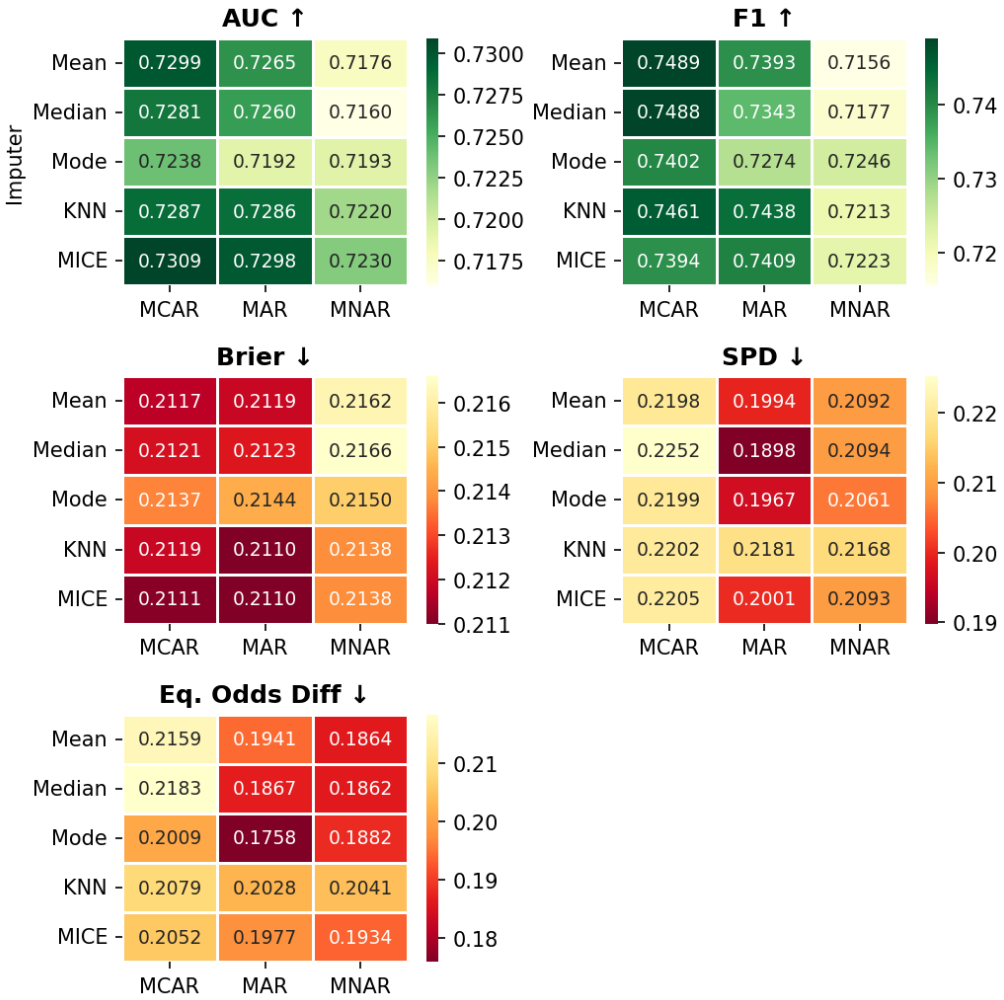


Figure 1. Performance metric heatmaps score. Color intensity indicates metric magnitude, with darker shades representing better performance (green), and better calibration (red).

exhibited degradation, with RMSE values increasing approximately threefold compared to MCAR and MAR. Mean maintained the best performance under MNAR (1218.31), but even this represents a $\approx$ 114% increase over its MAR performance. Mode imputation showed the highest RMSE under MNAR (2453.44), indicating severe point-wise reconstruction failures despite its superior downstream predictive performance reported in Section 4.2.

Table 3 presents mean KL-divergence values for the same configurations. Under MCAR, Mode achieved the lowest KL-divergence (0.030), indicating superior distributional preservation despite its high RMSE. KNN (0.035) and MICE (0.041) showed slightly higher but still excellent distributional fidelity. Mean imputation exhibited the highest KL-divergence (0.091), reflecting the artificial peaks introduced at central tendency values visible in Figure 2. Under MAR, all methods achieved similarity KL-divergence values, with Mode (0.029) and KNN (0.032) showing high distributional preservation. Under MNAR, however, KL-divergence increased across all methods - a 27-87$\times$ increase compared to MCAR and MAR. This increase indicates that MNAR

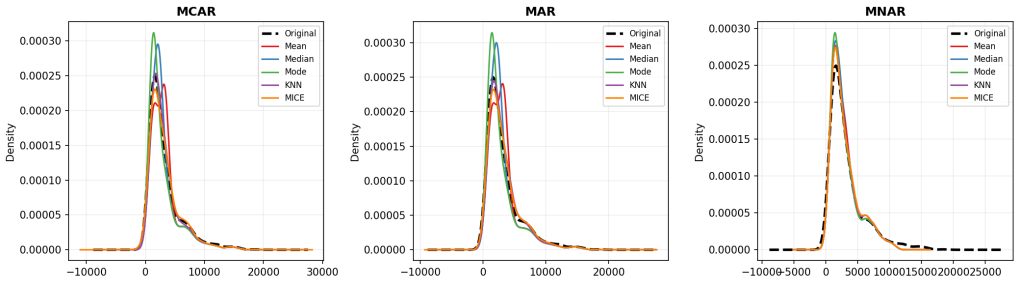


Figure 2. Kernel density estimates comparing the distributions of variable "credit_amount" after imputation against original complete data. Each panel shows the representative distributions, illustrating how imputation methods preserve or distort distributional characteristics.

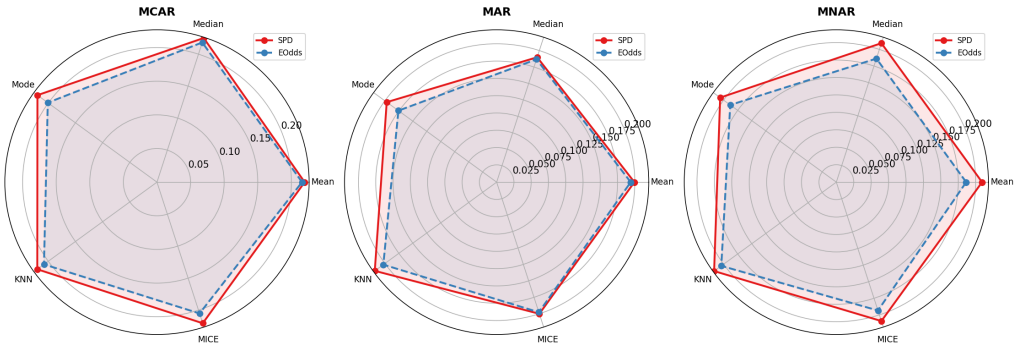


Figure 3. Radar plots displaying Statistical Parity Difference and Equalized Odds Difference across all configurations. Each axis represents a fairness metric, with distance from center indicating magnitude of disparity. Smaller enclosed areas indicate better overall fairness.

Table 2. Average RMSE across. Lower values indicate better point-to-point reconstruction accuracy. MNAR causes degradation with an approximately 2$\times$ increase compared to MCAR/MAR. The MICE imputor showed the best results.

Imputer	MCAR	MAR	MNAR
Mean	550.92	571.01	1218.31
Median	572.68	594.72	2258.90
Mode	666.29	686.5	2453.44
KNN	483.69	566.00	1945.83
MICE	438.88	459.59	1826.42

fundamentally breaks distributional fidelity regardless of imputation method sophistication. KNN achieved the lowest KL-divergence under MNAR (2.244), followed closely by MICE (2.252), while Mean and Median showed the highest distortion (≈ 2.55).

Table 3. Mean KL-divergence $D_{\mathsf{KL}}(P_{\mathsf{orig}} \| P_{\mathsf{imp}})$ across five numerical features. Lower values indicate better distributional preservation. MNAR causes distortion with KL increasing from $<$0.03 to $\sim$2.5

Imputer	MCAR	MAR	MNAR
Mean	0.091	0.094	2.545
Median	0.053	0.057	2.546
Mode	0.030	0.029	2.547
KNN	0.035	0.032	2.244
MICE	0.041	0.046	2.252

Figure 4 presents a comprehensive comparison of RMSE and KL-divergence across all five imputation methods in a 2×3 grid layout. Each panel shows per-column RMSE (top row) and KL-divergence (bottom row) for one mechanism. This visualization reveals the RMSE-KL trade-off: Mode imputation consistently achieves the lowest KL-divergence (best distributional preservation) but the highest RMSE (worst point-wise accuracy) under MCAR and MAR. Conversely, MICE achieves the lowest RMSE but higher KL-divergence than Mode. KNN occupies a middle ground, balancing both objectives. Under MNAR, the trade-off collapses as both metrics degrade catastrophically across all methods, though relative rankings persist with MICE/KNN showing the best performance on both metrics despite absolute values being unacceptably high.

The imputation quality results reveal three critical findings. First, MNAR causes degradation in both point-wise accuracy (RMSE increases 3$\times$) and distributional fidelity (KL increases 150-500×), confirming that non-ignorable missingness fundamentally breaks standard imputation assumptions. Second, under ignorable mechanisms (MCAR/MAR), a clear trade-off exists between RMSE and KL-divergence: methods optimizing point-wise accuracy (MICE, KNN) sacrifice some distributional fidelity, while methods preserving distributions (Mode) make larger point-wise errors. Third, the choice between RMSE and KL optimization depends on downstream task requirements - predictive modeling may prioritize RMSE, while distributional analysis or statistical inference requires low KL-divergence. These findings complement the predictive performance results in Section 4.2, demonstrating that imputation quality assessment requires multiple perspectives beyond downstream classification accuracy.

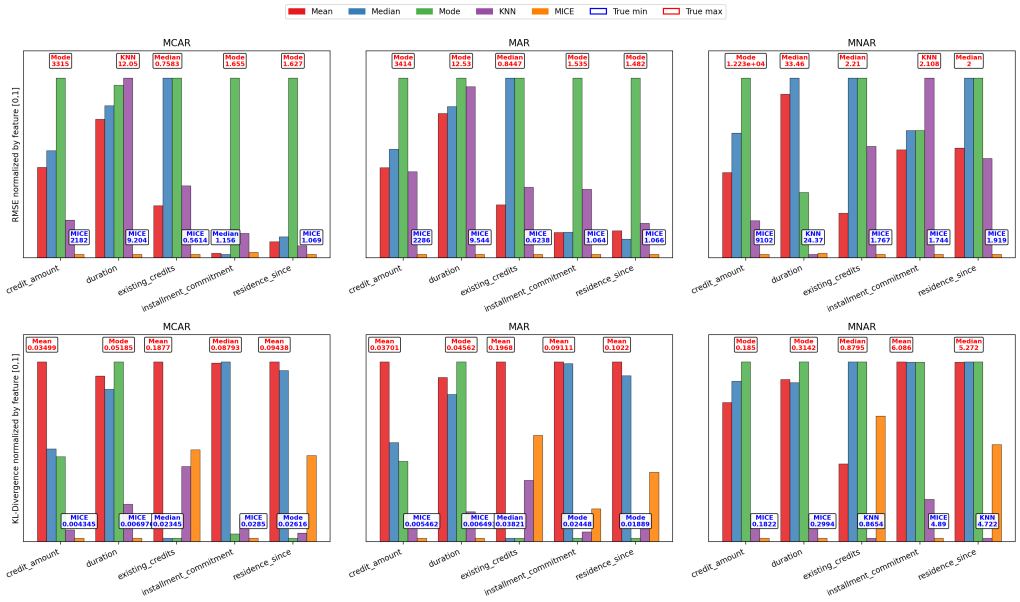


Figure 4. Combined RMSE and KL-divergence comparison across mechanisms (columns) and metrics (rows). Bars are grouped by imputation method, revealing the RMSE-KL trade-off under MCAR/MAR and the degradation under MNAR.

5. Discussion

Our findings confirm that imputation performance depends on the missingness mechanism, and that method selection must balance predictive accuracy and fairness. MICE excelled under MCAR and MAR, while Mode unexpectedly outperformed more complex methods under MNAR. These results challenge the notion of a universal best imputer and highlight the need for mechanism-aware strategies in credit risk modeling.

Mechanism-Specific Performance

MICE delivered superior results under ignorable mechanisms by capturing multivariate relationships, while Mode proved more robust under MNAR. Mean and Median consistently underperformed, particularly when missingness correlated with unobserved values.

Imputation Quality and Distributional Fidelity

RMSE and KL-divergence analyses reveal three insights: (i) MNAR causes severe degradation in both point-wise accuracy and distributional fidelity, (ii) under MCAR and MAR, a trade-off exists between RMSE (favoring MICE/KNN) and KL-divergence (favoring Mode), and (iii) downstream performance under MNAR does not align with imputation quality, as Mode achieved the best predictive results despite poor RMSE and KL scores.

Fairness-Performance Trade-offs

Fairness outcomes varied across mechanisms. MICE achieved the best fairness under MAR, while MNAR configurations generally reduced disparities across methods. Mode maintained intermediate fairness levels, underscoring the trade-off between accuracy and equity.

Practical Implications

The results suggest that diagnosing the mechanism of missing data before selecting

the imputer is a good practice. In MCAR/MAR data cases, MICE proved to be the best for predictive modeling, Mode for distributional analysis, and KNN offers a balance between the two. In MNAR data cases, standard imputators fail, and specialized methods or full case analysis may be necessary. Evaluation frameworks should incorporate performance and fairness metrics, as optimizing only one dimension can lead to unacceptable results.

Limitations

The results are based on the German credit dataset and logistic regression, which limits its generalization to larger and more complex portfolios and to alternative classifiers. Simulation with missing data may not fully capture all real-world patterns, and stock valuation was restricted to two metrics.

6. Conclusion

This study shows that missingness mechanism directly shapes imputation performance in credit risk modeling, and that imputer choice must balance accuracy and fairness. Evaluation of five methods across three mechanisms revealed distinct hierarchies, challenging the notion of a universal best approach.

This work contributes by: (i) benchmarking imputation–fairness interactions on the German Credit Dataset; (ii) showing that mechanism-aware imputation improves outcomes compared to mechanism-agnostic strategies; (iii) offering guidance - diagnose missingness before imputer selection, use MICE under MCAR/MAR, consider Mode under MNAR, and evaluate both accuracy and fairness; and (iv) establishing that no single method is optimal across mechanisms.

Future research should validate findings across diverse datasets, test advanced classifiers, investigate real-world missingness in production systems, develop MNAR-specific imputers, and explore fairness-aware approaches that jointly optimize accuracy and equity.

References

Caton, S., Malisetty, S., and Haas, C. (2022). Impact of imputation strategies on fairness in machine learning. *Journal of Artificial Intelligence Research*, 74:1011–1035.

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, 28(4):857–874. Retail Credit Risk Management and Measurement.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Hofmann, H. (1994). German credit data.

Khan, F. A., Herasymuk, D., Protsiv, N., and Stoyanovich, J. (2025). Still more shades of null: An evaluation suite for responsible missing value imputation.

Kwon, S., Jang, J., and Kim, C. O. (2025). Credit scoring using multi-task siamese neural network for improving prediction performance and stability. *Expert Systems with Applications*, 259:125327.

Leite, L. R. (2024). Comparative study of imputation methods for credit card delinquency prediction. Master's thesis, USP - Universidade de São Paulo, São Carlos, São Paulo, Brazil.

Li, B., Gao, R., Chen, Z., and Wang, C. D. (2025). Inference of rejection in credit scoring using the nrmpi method: A non-random missing data approach. In *Digital Finance*, pages 75–95, Singapore. Springer Nature Singapore.

Nastoska, A., Jancheska, B., Rizinski, M., and Trajanov, D. (2025). Evaluating trustworthiness in ai: Risks, metrics, and applications across industries. *Electronics*, 14(13).

Nogueira, I. (2025). Balancing accuracy and explainability in credit scoring: a comparative study of machine learning models and imputation methods. Master's thesis, Orebro University School of Business.

Putri, K. A., Hervani, N. M., Purwani, F., et al. (2025). Overcoming bias from missing values in microcredit data by comparing mice and imputation methods. *International Journal of Multidisciplinary Reseach*, 1(5):371–380.

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

Sue, K.-L., Tsai, C.-F., and Tsau, H.-M. (2024). Missing value imputation and the effect of feature normalisation on financial distress prediction. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(8):1467–1483.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Zhang, M., Xiao, Z., Wang, N., Fu, X., Cheng, Y., Wang, W., and Zhao, J. (2026). Jump: Jointly utilizing missingness pattern in tabular data. In *ICLR 2026 International Conference on Learning Representations*, Rio de Janeiro.

Zhang, Y. and Long, Q. (2021). Fairness in missing data imputation.