STABILITY
Journal of Management & Business
Vol. 9 No. 1 Years 2026
http://journal.upgris.ac.id/index.php/stability

FINANCIAL RISK MANAGEMENT: EVALUATING DATA IMPUTATION STRATEGIES FOR ACCURATE CREDITWORTHINESS PREDICTIONS

Rina Diah Wahyuningsih [1], Hasan Aminda Syafrudin [2]

Departement of Digital Business, Economics Business and Social Humanities Faculty, Universitas Telogorejo Semarang, Indonesia[1,2]

Article Information

Article History:
Submission : Mei 2026
Accepted : Juni 2026
Published : Juli 2026

Keywords:
Creditworthiness;
Data Imputation;
K-Nearest Neighbor;
Missing Value;
Risk Mitigation.

INDEXED IN

SINTA - Science and Technology Index
Dimensions
Google Scholar
ReseachGate
Garuda

Abstract

Predicting creditworthiness is crucial for mitigating loan risks, but missing data and extreme outliers often compromise model accuracy. This study evaluates the impact of data imputation quality on financial credit predictions using the K-Nearest Neighbor algorithm. Utilizing a Finance Loan Approval dataset of 614 records, this research compared four imputation methods: Mean, Median, Linear Interpolation, and KNN Imputer. The results demonstrate that outlier-robust imputation methods significantly enhance predictive integrity. Median and Linear Interpolation achieved the highest accuracy of 85.37% and an F1-Score of 90.32% by maintaining the spatial distances essential for the algorithm. Conversely, Mean imputation produced the lowest accuracy of 83.71% due to distortions from extreme income anomalies. Furthermore, the evaluation exposed algorithmic bias driven by class imbalance. The model achieved a near-perfect Recall of 99% but struggled with a lower Precision of 83%, indicating a high rate of false positives. In conclusion, while robust imputation strategies restore data structure and optimize accuracy, they cannot resolve class imbalance biases. Future scoring systems must integrate outlier-resistant imputation with advanced resampling techniques to improve precision and mitigate financial risks.

Correspondence Address :
Institusion Address
E-mail: (The email written is the email the author corresponds with the editor)

DOI : http://dx.doi.org/10.26877/vw6bnh14

OPEN ACCESS
ISSN :2621-850X
E-ISSN : 2621-9565

INTRODUCTION

The financial industry currently faces a critical challenge in the form of a high non-performing loan ratio, which threatens institutional asset stability and liquidity (Kar & Alam, 2025). This phenomenon often stems from system failures in accurately assessing prospective debtors risk profiles due to flawed or incomplete historical customer information (Clintworth, 2020). The massive digitalization of financial services has triggered a surge in credit application data volume, but simultaneously creates information asymmetry when customers fail to provide crucial data completely (Tian & Wei, 2025). This research must be conducted immediately because managerial decisions based on incomplete financial data will produce biased algorithmic outputs, ultimately leading to misdirected credit approvals and direct financial losses for the institution.

The selection of the historical financial loan application dataset as the research object is based on the characteristics of real-world banking data, which is highly susceptible to anomalies and the incompleteness of key attributes, such as credit history and income (Wang & Chen, 2025). Unlike other research objects like medical data or digital images whose distributions tend to be more even, credit approval data possesses two fundamental differences: a high rate of outliers due to financial disparities among customers, and the presence of class imbalance where the number of approved credits naturally always exceeds the rejected ones (Ndama et al., 2025). These extreme characteristics make the financial dataset the most ideal and challenging testing ground to measure the robustness of a data quality improvement method before its implementation in a credit scoring system.

Financial datasets rarely exhibit Missing Completely at Random mechanisms, where data loss is entirely independent of any variables . Instead, credit applications predominantly display Missing at Random patterns, where missing values correlate with other observed applicant features, or Missing Not at Random patterns, where applicants intentionally withhold disadvantageous financial information. This theoretical mechanism fundamentally explains why traditional Mean imputation often fails. Mean imputation mathematically assumes both a normal distribution and a strict Missing Completely at Random mechanism, causing it to severely distort the highly skewed structures inherent in financial data. In contrast, Median imputation offers a superior theoretical advantage for skewed Missing Not at Random scenarios because its central tendency calculation explicitly resists extreme financial outliers. Furthermore, the K-Nearest Neighbor Imputer is theoretically optimal for Missing at Random conditions. By leveraging spatial correlation and local geometry, the K-Nearest Neighbor algorithm reconstructs missing attributes by borrowing information from structurally similar observed profiles, directly addressing the interdependencies among variables.

Within the context of this predictive architecture, there is a strong interrelationship between the quality of data imputation as the independent variable and the accuracy of creditworthiness prediction as the dependent variable (Chakraborty & Ranjan, 2024).

Data imputation quality, which encompasses techniques for filling information gaps using Mean, Median, Linear Interpolation, and K-Nearest Neighbor (KNN) Imputer methods, directly determines the integrity and structure of the training data. This data integrity subsequently affects the dependent variable, where data restored with the appropriate method will enhance the classification algorithm's ability to recognize risk patterns precisely, evidenced by improvements in Accuracy, Precision, Recall, and F1-Score metrics. Conversely, an incorrect imputation method will distort the relationships between financial variables and cause a drastic decline in predictive performance (Çetin & Ahmed, 2023; Padimi et al., 2022).

Although creditworthiness classification has been extensively studied, there remains a significant research gap regarding the evaluation of pre-processing data quality, particularly under the constraints of class imbalance. Previous research by Chakraborty & Ranjan (Chakraborty & Ranjan, 2024) focused on optimizing the parameters of the KNN classification algorithm in the banking sector but neglected systematic interventions for missing data. Furthermore, the most recent study by Khan (Khan et al., 2024) applied advanced imputation algorithms to restore financial data, but the testing was only conducted on a proportionately balanced synthetic dataset, thus failing to capture the class imbalance phenomenon common in the real industry. Expanding on this, a recent investigation by Hartomo (Hartomo et al., 2025) highlighted the vulnerability of minority classes—such as rejected loan applications—to data distortion during the preprocessing phase, revealing that inappropriate missing value handling exacerbates model bias and causes the algorithm to over-predict the majority class. The primary gap in the majority of existing studies is the lack of comprehensive comparison between computationally efficient basic imputation methods like Mean, Median, Linear Interpolation with complex algorithmic imputation methods KNN-Imputer, specifically within an imbalanced financial data environment where imputation is treated as a critical independent variable rather than a minor step.

Based on the summary of these research gaps, the novelty of this study lies in the performance comparison of simple statistical imputation methods against proximity-based imputation on a financial dataset distorted by class imbalance. This research provides strategic benefits for risk managers and financial information system practitioners in the form of validated data governance guidelines to handle missing values without triggering algorithmic bias. Therefore, the main objective to be achieved from this research is to empirically evaluate the impact of various data imputation methods quality on the KNN algorithms accuracy in predicting creditworthiness, as well as to formulate the most resilient data recovery technique against financial outliers.

METHOD

This study employs a quantitative experimental research design to comprehensively evaluate the impact of various missing data handling techniques on the performance of predictive models. The primary research instrument utilized in this study is a Python version 3.10 computational environment running on a macOS operating system, specifically leveraging the Scikit-learn version 1.2.2 machine learning library to execute the data preprocessing, imputation, and classification algorithms. To ensure computational reproducibility across all experimental stages, a global random seed of 42 was strictly maintained. Recent literature frequently highlights Scikit-learn as a highly reliable and computationally efficient framework for reproducible machine learning research (Murzagaliyev et al., 2024). The data source is secondary data obtained from the Kaggle online public repository, specifically the Finance Loan Approval Prediction Data published by krishnaraj30. The dataset was accessed and virtually processed in 2026, representing a globally recognized financial dataset structure that is frequently utilized for credit risk algorithmic benchmarking (Jain et al., 2025).

This study directly utilizes the publicly available Kaggle benchmark dataset. This approach ensures that the dataset retains its authentic real-world financial anomalies, notably the presence of random missing values in crucial columns such as credit history and loan amount, alongside a highly imbalanced target class distribution. Methodological studies in data science validate the use of complete benchmark datasets to stress-test algorithms against known data quality issues without introducing sampling bias (Pandey et al., 2026). By utilizing the entirety of the provided data, the exact record size utilized in this research is 614 applicant profiles. This volume is highly appropriate and statistically sufficient for training, validating, and testing the chosen machine learning classification model without risking severe overfitting, as supported by empirical guidelines for basic algorithmic validation (Parveen & Thangaraju, 2024).

The analysis technique involves a series of systematic computational and statistical stages. Initially, the raw data undergoes a preprocessing phase, which includes categorical manual encoding and Min-Max scaling to normalize the numerical feature ranges to prevent distance bias. Ensuring all features share an identical scale is a mandatory theoretical requirement for distance-based classifiers like the K-Nearest Neighbor algorithm to function correctly (Kumar et al., 2025). Subsequently, the dataset is stratified and split into 80% training data and 20% testing data to prevent data leakage. This specific partition ratio is a widely accepted standard in machine learning to balance model training exposure with robust unseen data validation (Orji et al., 2022). The core experiment is then conducted by applying four distinct data imputation scenarios to handle the missing values in the training set: Mean imputation, Median imputation, Linear Interpolation, and the proximity-based KNN Imputer. Following the imputation phase, the reconstructed datasets are processed using the KNN classification algorithm. The predictive performance of each imputation scenario is then quantitatively evaluated and compared

using a Confusion Matrix, specifically calculating the Accuracy, Precision, Recall, and F1-Score metrics to determine the most robust approach for mitigating financing risks. Incorporating comprehensive metrics like the F1-Score alongside Accuracy is crucial for providing an unbiased evaluation, especially when dealing with imbalanced financial data (Tran & Tham, 2025).

To ensure clarity in measurement and algorithmic processing, the variables involved in this study must be operationally defined. The independent variable in this research is the Data Imputation Quality, which represents the methodological intervention applied to the missing data. The dependent variable is the Prediction Accuracy, representing the creditworthiness decision output. Furthermore, several predictor variables extracted from the applicants profile are utilized in the distance calculation of the classification process. For ease of reading and structured comprehension, the detailed operational definitions of these interconnected variables are centralized and systematically presented in Table 1.

Table 1. Operational Definitions and Measurement Scales of Research Variables

Variable	Variable Type	Operational Definition	Data Scale
Data Imputation Quality	Independent	The specific mathematical or algorithmic approach used to estimate and fill missing values within the dataset (Mean, Median, Linear Interpolation, KNN Imputer).	Categorical (Nominal)
Prediction Accuracy	Dependent	The model's ability to correctly classify an applicant's creditworthiness, measured through evaluation metrics (Accuracy, Precision, Recall, F1-Score).	Numeric (Ratio)
Loan Status	Target	The final decision of the	Categorical (Binary)

		financial institution regarding the loan application, categorized as approved (1) or rejected (0).	
Credit History	Predictor	A record indicating whether the applicant has systematically repaid previous debts, serving as a primary indicator of default risk.	Categorical (Binary)
Applicant Income	Predictor	The gross monthly income of the primary applicant, used to assess financial capacity.	Numeric (Ratio)
Loan Amount	Predictor	The total principal amount of money requested by the applicant for financing.	Numeric (Ratio)

By systematically executing this experimental framework, the research guarantees a rigorous and objective evaluation of how data imputation quality impacts predictive models. The application of these robust validation metrics on an authentic, imbalanced financial dataset ensures that the resulting analysis produces valid and actionable insights for mitigating credit risks. These methodological steps serve as the foundational basis for formulating the strategic managerial recommendations elaborated in the subsequent results and discussion section.

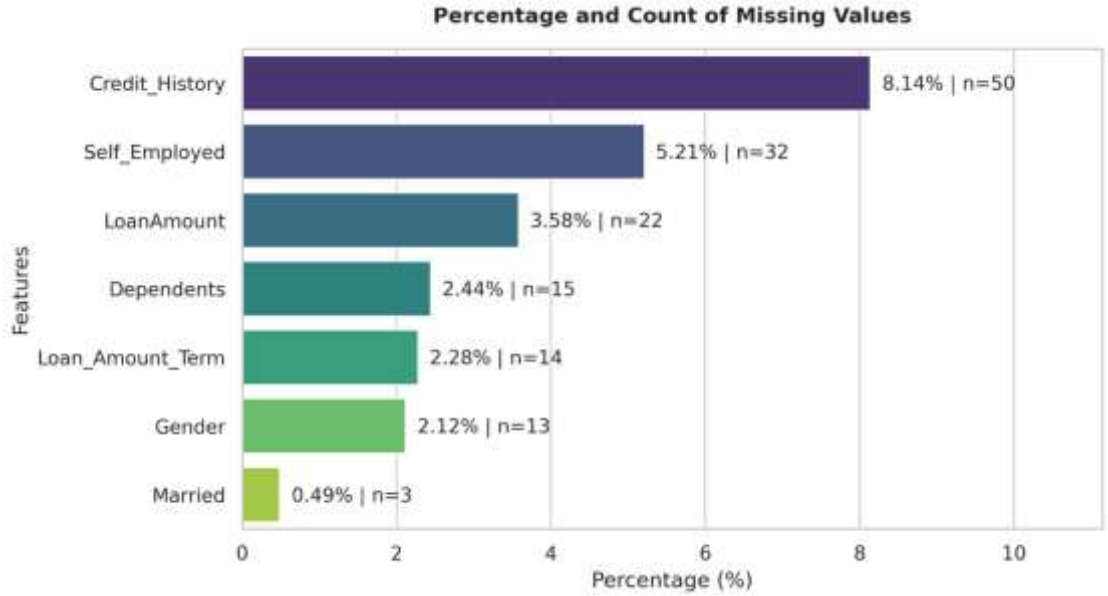
RESULTS AND DISCUSSION

Data Preprocessing and Distribution

The experiment utilized the Finance Loan Approval dataset consisting of applicant records. Initial exploratory data analysis revealed that missing values were widely distributed across seven distinct financial and demographic attributes. As illustrated in Figure 1, the Credit_History feature suffered the most severe data loss at 8.14 percent, which accounts for 50 missing observations. This was followed significantly

by Self_Employed at 5.21 percent with 32 missing observations, and LoanAmount at 3.58 percent with 22 missing observations. Other features such as Dependents, Loan_Amount_Term, Gender, and Married also exhibited minor missing values ranging from 0.49 percent to 2.44 percent, corresponding to between 3 and 15 missing observations. Prior to imputation, categorical variables were transformed using label encoding, and all numerical features were normalized using Min-Max Scaling to ensure that the KNN algorithm distance calculations were not dominated by variables with larger scales. This normalization step aligns with established machine learning principles, which assert that distance-based algorithms are highly sensitive to unscaled feature magnitudes, making normalization a mandatory prerequisite to prevent biased spatial estimations (Dornigg, 2021). Furthermore, the exploratory phase revealed a severe class imbalance in the target variable loanstatus, a common phenomenon in financial datasets that poses a distinct challenge as it creates a mathematical bias favoring the majority class. Previous studies on credit scoring risk emphasize that such imbalance often misleads standard evaluation metrics, causing models to aggressively over-predict the majority class while failing to detect high-risk minority profiles (Shih et al., 2022).

Figure 1. Percentage Distribution of Missing Values Across Financial Dataset Features



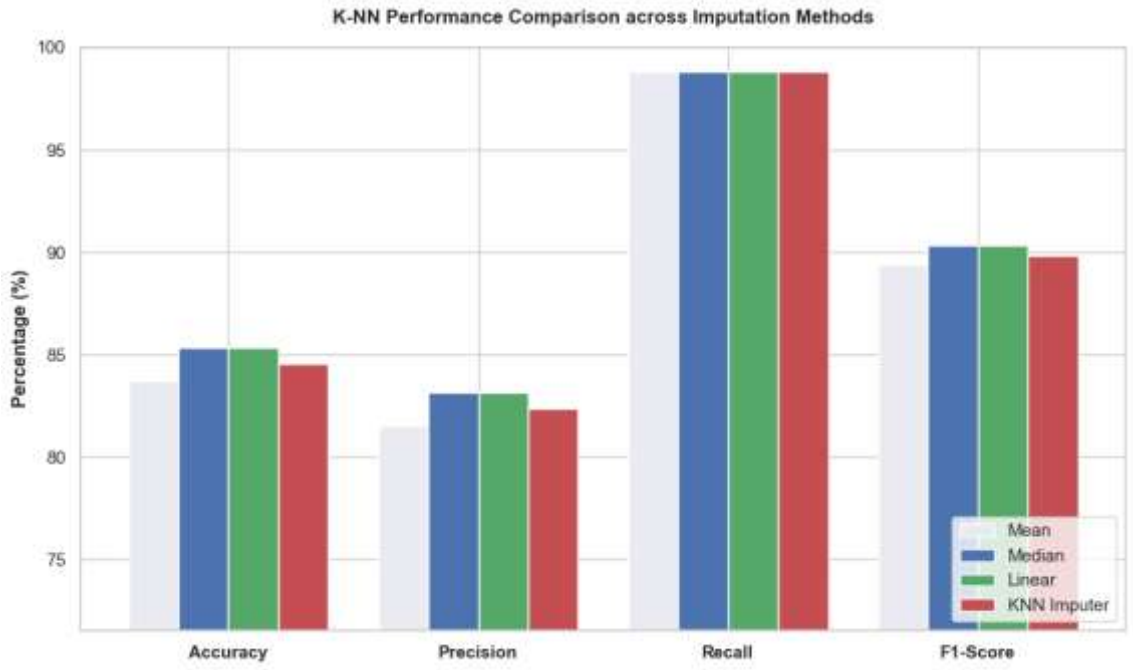
To restore the missing data distributed across the features shown in Figure 1, four distinct imputation scenarios were executed independently: Mean, Median, Linear Interpolation, and KNN Imputer. The selection of these specific strategies is grounded in recent comparative literature investigating data restoration methodologies for financial modeling (Yu et al., 2022). The Mean imputation replaced missing values with the statistical average, while the Median approach utilized the middle value of the sorted data. Linear Interpolation estimated missing points by drawing a straight line between adjacent data points. Lastly, the KNN Imputer dynamically filled gaps based on the mean values of the nearest data points in the multidimensional feature space. Existing research indicates that while traditional univariate methods like mean or median offer computational simplicity, advanced multivariate techniques such as the KNN Imputer are

theoretically better equipped to preserve the original variance and spatial geometry in complex datasets (Lu et al., 2025).

KNN Classification Performance

Following the completion of the data restoration phase, the KNN classification model was systematically trained and evaluated across the four distinct imputation scenarios. To rigorously assess the predictive capabilities and the behavioral tendencies of the algorithm, the evaluation utilized four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. The comparative empirical results of these metrics are comprehensively visualized in Figure 2.

Figure 2. Performance Comparison of K-Nearest Neighbor Algorithm Across Imputation Methods



In terms of overall Accuracy—which measures the proportion of correctly classified instances out of the total predictions—the data reveals a distinct performance hierarchy influenced by the imputation techniques. Both the Median and Linear Interpolation methods demonstrated superior and identical predictive stability, yielding the highest accuracy rates at approximately 85.37 percent. The KNN Imputer scenario closely followed with an accuracy of 84.55 percent. Conversely, the Mean imputation scenario recorded the lowest overall accuracy of 83.71 percent. This trajectory confirms that for a distance-based algorithm like K-NN, introducing mean-based values into a dataset heavy with outliers distorts the multidimensional spatial distances, thereby degrading general predictive correctness.

A more granular analysis of Precision and Recall unveils critical insights into the model's asymmetric learning behavior, heavily driven by the dataset's class imbalance. As depicted in Figure 2, the Recall metric reached exceptionally high levels across all scenarios. The Median, Linear, and KNN Imputer methods achieved near-perfect Recall scores of approximately 99 percent, with the Mean method slightly lagging but still

remarkably high. In a financial context, this indicates that the model is highly sensitive to the majority class; it successfully identifies and approves almost every actual creditworthy applicant as True Positives with minimal False Negatives.

However, this aggressive approval rate comes at the direct expense of Precision. The Precision scores plateaued at a significantly lower threshold, peaking at roughly 83 percent for the Median and Linear methods, and dropping to approximately 81.5 percent for the Mean approach. This stark disparity between Recall and Precision suggests an algorithmic bias where the K-NN model over-predicts the "Approved" class. Consequently, while the model rarely rejects a good applicant, it struggles to accurately isolate non-creditworthy applicants, leading to a substantial rate of False Positives, meaning risky debtors are classified as eligible.

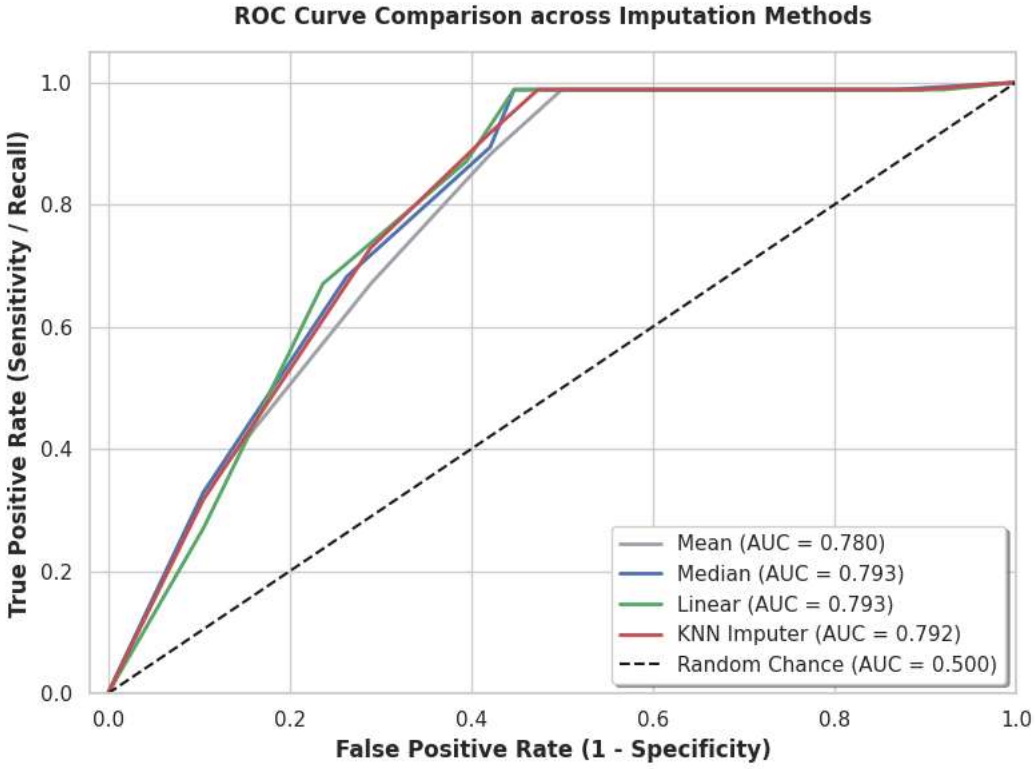
Table 2. Confusion Matrix Results of K-Nearest Neighbor Across Imputation Methods

Data Imputation Technique	TP	FP	TN	FN
Mean	84	19	19	1
Median	84	17	21	1
Interpolation	84	17	21	1
KNN Imputer	84	18	20	1

The exact distribution of these absolute predictions is detailed in the Confusion Matrix provided in Table 2. all imputation methods successfully identified 84 True Positives while only producing 1 False Negative. The divergence in performance stems entirely from the False Positives, where the Mean technique yielded the highest error with 19 misclassifications, compared to the optimal 17 misclassifications achieved by the Median and Linear Interpolation techniques.

To synthesize the trade-off between Precision and Recall, the F1-Score was evaluated as the harmonic mean of both metrics. Consistent with the primary findings, the Median and Linear imputations provided the most balanced and robust predictive performance, achieving the highest F1-Scores of approximately 90.32 percent. The KNN Imputer secured a slightly lower score of 89.8 percent, while the Mean method yielded the lowest harmonic balance of 89.4 percent.

Figure 3. Receiver Operating Characteristic Curve Comparison of Imputation Methods



To further evaluate the model's diagnostic ability to distinguish between creditworthy and risky applicants across different probability thresholds, an Area Under the Receiver Operating Characteristic Curve analysis was conducted. The resulting visual, depicted in Figure 3, corroborates the metric findings. Both the Median and Linear Interpolation methods achieved the highest Area Under the Curve score of 0.793. The KNN Imputer followed closely with a score of 0.792, while the Mean method lagged at 0.780. This curve visually demonstrates that median and interpolation techniques provide a marginally superior separation capability despite the overarching class imbalance.

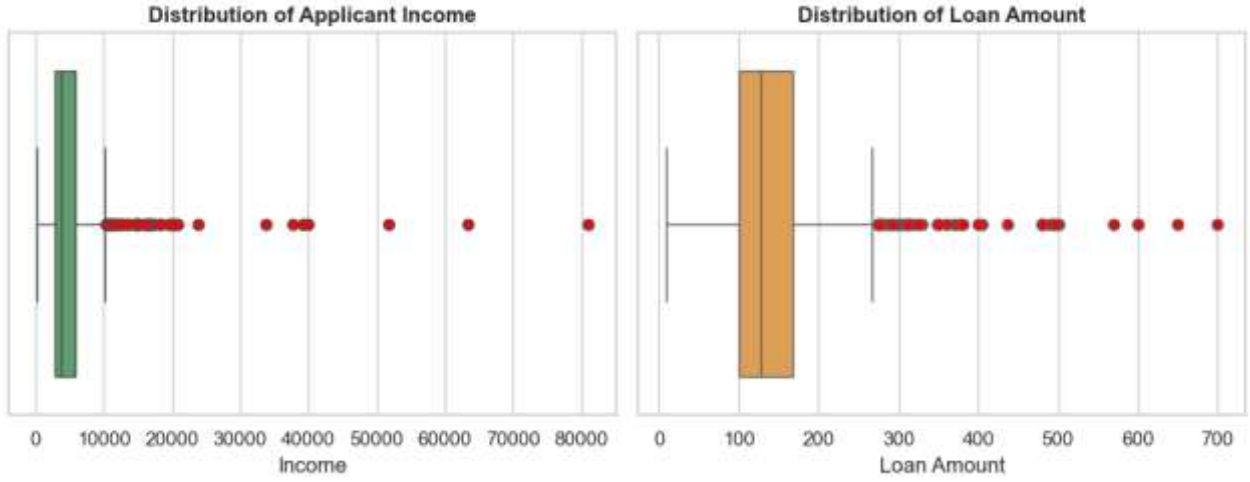
Ultimately, the classification performance data strongly substantiates that non-parametric and outlier-resistant data restorations—specifically Median and Linear Interpolation—provide the most mathematically sound foundation for distance-based modeling in skewed financial datasets. However, the pronounced gap between Recall and Precision explicitly highlights that while optimal imputation improves spatial integrity, it cannot independently neutralize the inherent biases caused by imbalanced target classes.

DISCUSSION

The empirical findings presented in the classification performance analysis reveal a profound intersection between data distribution characteristics, imputation strategies, and the fundamental mechanics of the KNN algorithm. As established, the Median and Linear Interpolation techniques consistently outperformed the Mean and KNN Imputer methods across primary evaluation metrics. To understand this performance gap, one must examine the spatial nature of the KNN algorithm alongside the structural integrity of the financial dataset. The KNN algorithm operates primarily on distance metrics, such

as Euclidean distance, to determine the proximity of data points in a multidimensional feature space. Therefore, the algorithm is hyper-sensitive to positional distortions. Various studies evaluating distance-based classifiers corroborate this finding, noting that unhandled outliers inherently poison the Euclidean space and drastically reduce classification reliability (Nandipati & Boddala, 2024). Figure 3 provides a critical visual explanation for why the Mean imputation underperformed. The boxplot distributions of ApplicantIncome and LoanAmount highlight the presence of extreme, right-skewed outliers—with applicant incomes irregularly stretching toward the 80,000 mark.

Figure 4. Boxplot Distribution Highlighting Extreme Outliers in Applicant Income and Loan Amount



When the Mean imputation calculates the average of such heavily skewed features, the resulting statistical value is artificially pulled toward the extreme outliers. Consequently, injecting these inflated mean values into the dataset distorts the spatial coordinates of the missing data points, pushing them away from their true natural clusters. This spatial distortion directly degrades the KNN model's ability to find the correct nearest neighbors, culminating in the lowest observed Accuracy of 83.71%. In contrast, the Median method remains statistically robust against extreme anomalies. By utilizing the 50th percentile, it ensures that the imputed values accurately reflect the densest concentration of the general population, thereby preserving the spatial geometry required by KNN for optimal classification. This empirical outcome aligns with recent comparative analyses in econometric machine learning, which emphasize that median-based or robust statistical restorations are strictly necessary when dealing with heavy-tailed financial distributions (Marevac et al., 2023). Conversely, inserting mean values into such skewed distributions creates artificial data points that do not represent any real applicant segment, thereby confusing the classifier (Khalili & Chulia, 2026).

Furthermore, the pronounced discrepancy between the near-perfect Recall of approximately 99% and the plateaued Precision of approximately 83% exposes a systemic vulnerability in the dataset: class imbalance. The model's tendency to

aggressively classify applicants as "Approved" demonstrates that the KNN algorithm has become biased toward the majority class. From a financial risk management perspective, this algorithmic behavior is highly problematic. While a high Recall ensures that legitimate clients are not unjustly denied credit, a low Precision indicates a significant rate of False Positives. In real-world banking operations, approving loans for non-creditworthy applicants directly inflates the risk of Non-Performing Loans, which can severely impact institutional liquidity. Recent literature on credit risk mitigation supports this assessment, arguing that evaluating models solely on overall accuracy or recall on imbalanced datasets creates a dangerous illusion of reliability (Zhang et al., 2025). Scholars emphasize that minimizing false positives through precision-focused metrics is critical to preventing cascading loan defaults (Ertuğrul, 2023).

Ultimately, the results confirm that selecting an outlier-resistant imputation method, such as the Median, is a mandatory foundational step that successfully optimizes spatial data integrity for distance-based algorithms. However, the discussion also firmly establishes that imputation alone is insufficient to resolve the behavioral biases induced by imbalanced data. To build a truly reliable credit scoring model, future iterations of this research must augment robust imputation techniques with algorithmic level interventions, such as resampling methods like Smote or cost-sensitive learning, to elevate Precision without sacrificing the integrity restored during preprocessing. Contemporary researchers have demonstrated that integrating hybrid sampling techniques like Smote alongside robust imputation significantly stabilizes the precision-recall trade-off in credit assessment frameworks (Wei, 2025).

CONCLUSIONS AND SUGGESTIONS

This study aimed to evaluate the impact of various data imputation techniques on the predictive performance of the KNN algorithm within the context of financial credit approval. The findings conclusively demonstrate that the choice of imputation method profoundly dictates the success of distance-based classification models. Techniques that are highly robust against extreme data outliers, specifically the Median and Linear Interpolation approaches, proved superior in preserving the original structural integrity of the dataset. Conversely, the Mean imputation method notably underperformed. The presence of extreme financial anomalies heavily skewed the calculated averages, which subsequently distorted the multidimensional spatial distances required by the KNN algorithm to accurately map and classify applicant profiles.

IMPLICATIONS AND LIMITATIONS

Theoretically, this research contributes to the machine learning preprocessing literature by reinforcing the principle that applying parametric imputation to non-parametric, outlier-heavy datasets actively damages the feature space in distance-based algorithms. Practically, for the financial sector and banking institutions, these findings emphasize that deploying outlier-resistant data restoration techniques is a critical

safeguard. By preventing the artificial inflation of applicant financial profiles during the preprocessing phase, institutions can build more reliable automated credit scoring systems, thereby ensuring that the spatial geometry of the data accurately reflects real-world economic conditions.

Despite the positive outcomes regarding spatial data restoration, this study possesses distinct limitations. Primarily, the research focused exclusively on resolving missing values and did not algorithmically intervene to correct the severe class imbalance inherently present in the financial dataset. As a result, the model developed an algorithmic bias toward the majority class; while it successfully identified creditworthy applicants, it struggled to accurately isolate non-creditworthy individuals, leading to a notable rate of false positives. Furthermore, the evaluation was limited to a standard KNN algorithm without extensive algorithmic modifications or hyperparameter tuning across diverse mathematical distance metrics. Based on the identified limitations, it is highly recommended that future research integrates the optimal robust imputation methods identified in this study with advanced data balancing techniques, such as synthetic minority oversampling or cost-sensitive learning frameworks. Addressing the class imbalance directly will help mitigate the false positive rates in credit prediction. Additionally, exploring the synergy between these data restoration strategies and more complex, ensemble-based machine learning algorithms could provide a more comprehensive and robust solution for handling missing data, extreme outliers, and imbalanced classes simultaneously in financial risk forecasting.

REFERENCES

Çetin, A. I., & Ahmed, S. E. (2023). Determinants of credit ratings and comparison of the rating prediction performances of machine learning algorithms. *E3S Web of Conferences*. https://pdfs.semanticscholar.org/d6fc/a45c0c08f3b91a7c9f4b548d230eef13ba62.pdf

Chakraborty, D. B., & Ranjan, R. (2024). Missing data imputation with contextual granules and AI-driven bankruptcy prediction. *2024 14th International …*. https://ieeexplore.ieee.org/abstract/document/10677843/

Clintworth, M. (2020). *Financial risk management in shipping investment, a machine learning approach*. stax.strath.ac.uk. https://stax.strath.ac.uk/concern/theses/8336h203f

Dornigg, T. (2021). *Credit Risk Modeling: Predicting Customer Loan Defaults with Machine Learning Models*. search.proquest.com. https://search.proquest.com/openview/fb1573d090f1f88de3fb50cbb68813ef/1?pq-origsite=gscholar&cbl=2026366&diss=y

Ertuğrul, S. (2023). *Customer transaction predictive modeling via machine learning algorithms*. search.proquest.com. https://search.proquest.com/openview/9db4a0358b28b01a9a4bf2b525f0f83f/1?pq-origsite=gscholar&cbl=2026366&diss=y

Hartomo, K. D., Arthur, C., & Nataliani, Y. (2025). A novel weighted loss tabtransformer integrating explainable ai for imbalanced credit risk datasets. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/10884750/

Jain, P., Vats, P., & Singh, S. (2025). Data-Driven Credit Risk Modeling for Corporate Debt Issuance: A Predictive Analytics Approach. *International Conference on Information and …*. https://doi.org/10.1007/978-3-032-20606-0_5

Kar, V., & Alam, K. M. (2025). Adaptive Ensemble Technique Based Robust Loan Risk Prediction for Financial Systems. *... on Computing Technologies &Data …*. https://ieeexplore.ieee.org/abstract/document/11158985/

Khalili, S., & Chulia, H. (2026). Data Imputation in Large Datasets: A Comparative Study of PCA and Machine Learning Approaches. *Computational Economics*. https://doi.org/10.1007/s10614-025-11201-x

Khan, F. A., Herasymuk, D., Protsiv, N., & ... (2024). Still More Shades of Null: An Evaluation Suite for Responsible Missing Value Imputation. *ArXiv Preprint ArXiv …*. https://arxiv.org/abs/2409.07510

Kumar, S., Singh, S. K., Singh, N. P., Sagu, A., & ... (2025). An Enhanced Credit Risk Classification Framework Using Hybrid Genetic Algorithm and Machine Learning Models on the German Credit Dataset. *Cureus Journals*. https://www.researchgate.net/profile/Susheel-Kumar-46/publication/396502762_An_Enhanced_Credit_Risk_Classification_Framework_Using_Hybrid_Genetic_Algorithm_and_Machine_Learning_Models_on_the_German_Credit_Dataset/links/68ef3bf302d6215259bbc815/An-Enhanced-Credit-Risk-Classification-Framework-Using-Hybrid-Genetic-Algorithm-and-Machine-Learning-Models-on-the-German-Credit-Dataset.pdf

Lu, J., Zhuo, S., Qiu, J., & Tang, Y. (2025). Toward accurate credit evaluation: An efficient imputation approach for financial data. *Data Science and Management*. https://www.sciencedirect.com/science/article/pii/S2666764925000281

Marevac, E., Patković, S., & Žunić, E. (2023). Decision-making AI for customer worthiness and viability. *2023 22nd International …*. https://ieeexplore.ieee.org/abstract/document/10094207/

Murzagaliyev, R., Sagidolla, B., & Kadyrov, S. (2024). Credit Score Prediction Through Hybrid Machine Learning Models. *UiTM International Conference …*. https://doi.org/10.1007/978-981-96-9350-4_12

Nandipati, V. S. S., & Boddala, L. V. (2024). *Credit card approval prediction: A comparative analysis between logistic regression, knn, decision trees, random forest, xgboost*.

Ndama, S., Ndama, O., & En-Naimi, E. M. (2025). A Comprehensive Evaluation of Data Balancing Techniques and Machine Learning Models for Credit Risk Assessment. *International Conference on Digital …*. https://doi.org/10.1007/978-3-032-06725-8_12

Orji, U. E., Ugwuishiwu, C. H., & ... (2022). Machine learning models for predicting bank loan eligibility. *2022 IEEE Nigeria …*. https://ieeexplore.ieee.org/abstract/document/9803172/

Padimi, V., Venkata, S. T., & Devarani, D. N. (2022). Applying machine learning techniques to maximize the performance of loan default prediction. *Journal of Neutrosophic and …*. https://www.researchgate.net/profile/Devarani-Devi-Ningombam/publication/360032011_Applying_Machine_Learning_Techniques_To_Maximize_The_Performance_of_Loan_Default_Prediction/links/625e40cb4173a21a0d1dabe5/Applying-Machine-Learning-Techniques-To-Maximize-The-Performance-of-Loan-Default-Prediction.pdf

Pandey, P., Himanshu, A. K., & Gola, K. K. (2026). Predictive Modeling of Credit Card Approvals Using Ensemble Learning Techniques. *2026 2nd International …*. https://ieeexplore.ieee.org/abstract/document/11469142/

Parveen, K. R., & Thangaraju, P. (2024). Enhanced credit scoring prediction using KNN-Z-score based logistic regression (KZ-LR) algorithm. *J. Electr. Syst*. https://www.researchgate.net/profile/Rizwana-Parveen-5/publication/387549642_Enhanced_Credit_Scoring_Prediction_Using_KNN-Z-Score_Based_Logistic_Regression_KZ-LR_Algorithm/links/677b72e3e74ca64e1f4ecfa8/Enhanced-Credit-Scoring-Prediction-Using-KNN-Z-Score-Based-Logistic-Regression-KZ-LR-Algorithm.pdf

Shih, D. H., Wu, T. W., Shih, P. Y., Lu, N. A., & Shih, M. H. (2022). A framework of global credit-scoring modeling using outlier detection and machine learning in a P2P lending platform. *Mathematics*. https://www.mdpi.com/2227-7390/10/13/2282

Tian, G., & Wei, M. (2025). Explainable Shapley Additive Explanations-Enhanced Stacked Ensemble for Digital Financial Risk Prediction. *2025 3rd International Conference on Data …*. https://ieeexplore.ieee.org/abstract/document/11168558/

Tran, D., & Tham, A. W. (2025). Accuracy comparison between feedforward neural network, support vector machine and boosting ensembles for financial risk evaluation. *Journal of Risk and Financial Management*. https://www.mdpi.com/1911-8074/18/4/215

Wang, S., & Chen, Y. (2025). Digital Financial Risk Early Warning Using Particle Swarm Optimization and Bidirectional Long Short-Term Memory. *2025 3rd International Conference on Data …*. https://ieeexplore.ieee.org/abstract/document/11168509/

Wei, H. (2025). SMOTE algorithm optimization and application in corporate credit risk prediction with diversification strategy consideration. *Scientific Reports*. https://www.nature.com/articles/s41598-025-09173-x

Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2022). Missing data preprocessing in credit classification: One-hot encoding or imputation? *Emerging Markets Finance and …*. https://doi.org/10.1080/1540496X.2020.1825935

Zhang, C., Zhou, W., Zhao, X., & Yang, S. (2025). A machine learning framework for missing and imbalanced data in marketing analytics. *Journal of Marketing Analytics*. https://doi.org/10.1057/s41270-025-00432-4