SMU Data Science Review

Volume 8
Number 2 *Summer 2024*

Article 5

2024

Enhancing Imputation Accuracy: A Multi-Faceted Approach for Missing Data in Chicago Arrest Records

Steve Bramhall
txbramhall@verizon.net

Jae Chung
Southern Methodist University, jaegunchung93@gmail.com

Nicholas Mueller
Southern Methodist University, nmueller111@yahoo.com

Enhancing Imputation Accuracy: A Multi-Faceted Approach for Missing Data in Chicago Arrest Records

Jae Chung[1], Nicholas Mueller[1], Steve Bramhall[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] Vistra Corporation, 1925 W John Carpenter Fwy, #300
Irving, TX 75063 USA
{jgchung, nmueller}@smu.edu
txbramhall@verizon.net}

Abstract. This paper introduces a novel approach to enhance the imputation process for missing data, utilizing crime records from Chicago with arrests as the target feature. Robust imputation techniques are crucial in the era of burgeoning datasets for generating reliable insights. Our core objective is to present an innovative method that improves imputation techniques, augmenting model performance and bolstering the reliability of analytical outcomes. Leveraging numeric crime data, we establish a Gradient Boosting (GBM) baseline model, then introduce ensemble methods including Random Forest and Decision Trees for further refinement. By systematically exploring multiple imputation processes, we establish a baseline for comparative analysis, enabling precise measurement of efficacy. Inspired by existing literature, our imputation process elevates performance metrics and provides actionable insights. This study addresses broader challenges in data imputation, particularly in crime data analysis in urban settings like Chicago. Throughout, we document our methodology, experimentation, and findings, highlighting the effectiveness of ensemble techniques coupled with GBM in addressing data imputation challenges. Our research aims to empower practitioners and researchers with enhanced decision-making capabilities and analytical prowess in data-rich environments.

1 Introduction

The safety of the public is a fundamental aspect of society. However, for certain communities, safety remains a distant luxury rather than a guaranteed norm. In recent years, Chicago has garnered widespread attention for the prevalent chaos and disorder that permeates its streets. Despite concerted efforts in other cities to curb crime rates and bolster public safety, Chicago has struggled to make significant strides. A comparative analysis of homicide rates between Chicago, New York City, and Los Angeles from 1990 to 2023 illustrates this stark disparity. While New York City witnessed a dramatic decline from 2,245 to 386 homicides and Los Angeles experienced a notable decrease from 991 to 327 homicides during this period, Chicago's reduction was comparatively modest, dropping from 849 to 617 homicides (Guthmann, 2024).

Centralized within the metropolitan area, approximately 42% of Chicago's total homicides are concentrated in the five most perilous and violent neighborhoods (Guthmann, 2024). However, this alarming statistic underscores a broader concern: more than half of all homicides occur in areas not traditionally identified as high-risk by law enforcement. Recognizing the imperative for community support, many cities are acknowledging the significance of predictive crime analytics in enhancing public awareness and fostering proactive strategies for crime prevention (Guthmann, 2024). To safeguard both the public and law enforcement personnel, it is imperative to adopt a proactive approach that hinges upon a comprehensive understanding of the prevailing circumstances, necessitating access to comprehensive datasets.

Despite strides made in understanding crime patterns, significant challenges persist in effectively predicting and preventing criminal activities. In light of Chicago's persistent struggles with crime reduction, there is a pressing need for innovative strategies that leverage advanced data analytics and predictive modeling. By harnessing the power of insights motivated by data, policymakers and law enforcement agencies can devise targeted interventions, allocate resources strategically, and engage communities in collaborative efforts to address underlying socio-economic factors contributing to crime. By fostering a holistic approach that integrates data-driven methodologies with community engagement initiatives, cities like Chicago can aspire to mitigate crime rates and cultivate safer environments conducive to the well-being of all residents.

Moreover, addressing the multifaceted challenges of crime requires a nuanced understanding of its root causes and contributing factors. Socio-economic disparities, lack of access to education and employment opportunities, systemic inequalities, and community disinvestment are among the underlying issues fueling crime in urban areas like Chicago. Therefore, any effective crime reduction strategy must encompass not only law enforcement measures but also comprehensive social interventions aimed at addressing these systemic inequities. Investments in education, job training programs, affordable housing initiatives, and community development projects are crucial components of a holistic approach to crime prevention and community safety.

Furthermore, enhancing public safety entails creating a foundation of trust and collaboration between law enforcement agencies and the communities they serve. Building strong partnerships with community members, including residents, businesses, faith-based organizations, and non-profit groups, fosters a sense of shared responsibility and collective ownership in addressing crime and improving neighborhood safety. Community policing initiatives, restorative justice programs, and proactive engagement efforts can help unify law enforcement and the communities they serve, fostering mutual respect, understanding, and cooperation.

In essence, the quest for safer communities demands an approach that goes beyond traditional law enforcement tactics. By leveraging data-driven insights, addressing socio-economic differences, and fostering community partnerships, cities like Chicago can work towards creating environments where safety is not merely a luxury but a fundamental right for all residents. Such concerted efforts not only can reduce crime

rates but also enhance the quality of life and well-being for individuals and communities.

In data science and analytics, the accurate handling of missing data stands as a matter of immediate attention and fundamental challenge with profound implications for the reliability and integrity of analytical outcomes across diverse domains. Traditional imputation methodologies, while effective in many respects, often grapple with the complexities that reside in instances of intricate missing data patterns or the presence of observable variables. These challenges show the critical need for a pioneering approach capable of addressing these limitations and establishing a new benchmark for imputation accuracy.

Our research will attempt to fill and bridge this gap by integrating established models alongside innovative components, with a focus on adaptability and resilience when faced with complex data structures. Leveraging advanced techniques such as tensor-based imputation and insights derived from shared missing data patterns, our aim is to transcend the challenges associated with these complex datasets. However, our ambition extends beyond simply refining the imputation method. Rather, we will venture to shift the imputation landscape into one that will offer a solution capable of setting new standards for accuracy in complex data environments.

For data scientists, data researchers, and data-driven decision makers, our study holds significance in elevating the foundational principles of data accuracy, thereby enhancing the robustness of analytical models. By critically examining the limitations of existing imputation methodologies and introducing innovative strategies grounded in scientific rigor, our research can serve as a catalyst for advancing data integrity and fostering trust in decision-making processes that are reliant on and currently bottle-necked due to incomplete datasets.

Guided by the scientific framework of our predecessors, our research is anchored in several overarching objectives as previously mentioned. Firstly, we aim to articulate the inherent limitations of prevailing imputation models, thereby facilitating a deeper understanding of the challenges posed by intricate missing data patterns. Secondly, through the synthesis of established methodologies with new innovative components rooted in statistical theory, we aspire to establish a novel approach for model performance through the strategic utilization of ensembling techniques. Lastly, our research seeks to deliver a pragmatic solution that not only introduces a flexible and robust imputation framework, but also underscores its empirical validation when faced with real-world scenarios.

Through pursuing these objectives, our research ventures to contribute to the advancement of imputation techniques and methodologies. Our hope is that this will empower stakeholders across various sectors with the scientific tools, insights and confidence necessary to navigate the complexities of modern data analysis with confidence and precision.

2 Literature Review

2.1 Background on Chicago Crime Data

The city of Chicago has long grappled with the challenge of reducing its crime rates, particularly homicides, which have not declined as rapidly as in other major cities like New York and Los Angeles (Guthmann, 2024). Several factors contribute to Chicago's unique crime landscape, including historical racial inequalities, concentrated poverty, and gang-related violence (Guthmann, 2024)(FBI, 2017). Chicago's violent crime issue is often concentrated in specific neighborhoods, with a small proportion of areas accounting for a significant portion of homicides (FBI, 2017). Efforts to address these challenges have been hindered by various factors, including staffing shortages within the Chicago Police Department and the release of violent offenders due to bail reform (Vallas, 2023).

The urgency of addressing Chicago's crime problem has prompted calls for immediate action from city officials (Vallas, 2023). However, concrete plans to curb violence have been lacking, leaving Chicagoans searching for viable solutions (Vallas, 2023). Immediate steps proposed to address the crisis include fully staffing the police department, recruiting retired officers and military veterans, and improving 911 response times (Vallas, 2023). Additionally, measures to hold perpetrators accountable, enhance safety on public transit, and involve public schools in crime prevention efforts have been recommended (Vallas, 2023). These proposals aim to address both the immediate challenges of crime control and the underlying systemic issues contributing to Chicago's crime epidemic.

In conclusion, addressing Chicago's crime crisis requires a multifaceted approach that combines immediate interventions with long-term strategies (Guthmann, 2024)(FBI, 2017)(Vallas, 2023). Efforts to improve public safety must address staffing shortages, enhance law enforcement capabilities, and hold offenders accountable while also addressing root causes such as poverty, inequality, and lack of educational opportunities. Only through comprehensive and coordinated action can Chicago hope to effectively combat its persistent crime problem and create safer communities for all residents.

2.2 Limitations in Present-day Imputation Methods

The complexities surrounding the handling of missing data in statistical analyses, particularly through methods like Multiple Imputation by Chained Equations (MICE) and ensembling, present several formidable challenges that underscore the intricate nature of the task at hand. Firstly, the process' inherent complexity is highlighted by the substantial difficulties in defining an appropriate imputation model, exacerbated by the demanding computational requirements needed to execute these procedures effectively (Gurtskaia et al, 2024). Secondly, the significant time investment required for imputation is particularly notable, especially in scenarios involving labor-intensive processes such as resampling data or managing datasets with a vast number of variables,

further complicating the imputation process (Gurtskaia et al., 2024). Lastly, the risk of overfitting poses a substantial challenge; given that methods like MICE and ensembling rely heavily on model specifications, they are particularly vulnerable to issues such as overfitting and convergence difficulties. This susceptibility is especially pronounced in cases involving multicollinearity and similar forms of instability, which can severely affect the reliability and validity of the imputation outcomes (Gurtskaia et al., 2024).

2.3 Imputation Tools Utilized in Approach

The integration of ensemble machine learning models, such as RandomForest and GradientBoosting, signifies major advancements in predictive analytics. These models bring together predictions from many weak learners to produce outcomes that are statistically more robust and accurate than those derived from any single model (Delgado-Panadero et al., 2023). This cooperation between the various imputation methodologies leverages the unique strengths of the said various models, thereby mitigating the risk of overfitting and bolstering the reliability and robustness of the predictions made.

In parallel, the Multiple Imputation by Chained Equations (MICE) technique offers a sophisticated approach to managing missing data within datasets. By generating multiple imputations for incomplete observations, MICE meticulously fills in missing values by iteratively modeling each variable with missing data as a function of other variables present in the dataset. This process not only aids in creating a more complete dataset for analysis but also ensures that the imputed values are plausible and reflective of the underlying data structure (Gurtskaia et al, 2024).

Moreover, stacking, a sophisticated ensemble technique, holds promise as an effective imputation tool in the context of missing data handling. Stacking involves training multiple diverse models on the available dataset and using the predictions of these models as features for a meta-model, which then produces the final imputed values. This approach capitalizes on the diversity of the base models, leveraging their collective predictive power to generate more accurate imputations (Van der Laan et al, 2007). Stacking has been shown to outperform individual imputation methods by capturing complementary information from different models, thereby reducing bias and variance in imputed values.

Furthermore, the concept of sharing pattern submodels, as proposed by (Stempfle et al, 2023), introduces a novel approach to maintaining the reliability of predictions, even in the absence of certain data values during test time. This method not only maintains or improves the predictive accuracy of the underlying pattern submodels but also enhances their interpretability by offering succinct explanations. The implementation of sparsity-inducing regularization within this framework ensures that parameter sharing leads to consistent and precise estimations, further contributing to the method's efficacy.

Lastly, logistic regression, a fundamental statistical modeling technique, can also be utilized as an imputation tool, particularly in scenarios involving categorical or binary variables. Logistic regression estimates the probability of a binary outcome based on one or more predictor variables, making it suitable for imputing missing categorical data. By modeling the relationship between the observed variables and the missing variable of interest, logistic regression can impute missing values based on the available information, taking into account the underlying patterns and associations in the data (Allison, P. D., 2002). Logistic regression imputation has been demonstrated to yield reliable results, especially when the missingness mechanism is related to the observed variables and can be adequately modeled.

Incorporating these diverse methodologies into the imputation process underscores a comprehensive strategy for addressing some of the most pervasive challenges in data analysis and machine learning, ultimately enhancing the accuracy and reliability of analytical outcomes.

2.4 Modeling Methods

Extreme Gradient Boosting (XGBoost) stands out as a cornerstone in the realm of machine learning, representing a significant leap forward in predictive analytics. Engineered to harness the power of gradient boosted decision trees, XGBoost offers a high-performance library renowned for its exceptional utility across diverse tasks. A pivotal feature of XGBoost lies in its integration of regularization techniques aimed at mitigating overfitting, thereby enhancing its efficacy in handling complex datasets (Gurtskaia et al, 2024). Noteworthy attributes including efficiency, scalability, and adaptability underline its prominence in the field, making it a preferred choice for various machine learning applications.

In complement to XGBoost, the adaptation of Random Forest with Predictive Mean Matching presents another robust modeling method. This technique augments the inherent accuracy and reliability of a single decision tree by aggregating predictions from multiple trees. Each tree within the ensemble is cultivated from a distinct random data subset, ensuring a uniform distribution throughout all trees. This approach not only enhances the model's generalization capabilities but also substantially reduces the risk of overfitting, thereby solidifying its robustness across diverse analytical scenarios (Gurtskaia et al., 2024).

Furthermore, LightGBM, a Gradient Boosting Decision Trees (GBDT) method, addresses the inefficiencies observed in traditional gradient boosting methods like XGBoost and pGBRT. Particularly adept in scenarios characterized by high-dimensional features and large datasets, LightGBM introduces innovative techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). GOSS selectively retains data instances with significant gradients, thereby reducing computation time required for estimating information gain. On the other hand, EFB bundles mutually exclusive features to decrease feature dimensionality, leveraging a greedy algorithm to approximate the optimal bundling efficiently. While EFB may

encounter challenges in pinpointing the optimal bundling, it effectively achieves a notable reduction in feature count without compromising split-point determination accuracy significantly (Ke et al, 2017). These techniques collectively enhance LightGBM's flexibility and adaptability to various datasets, making it a valuable tool for accelerating the training process while maintaining competitive accuracy levels.

Linear regression stands as a fundamental yet powerful modeling technique in predictive analytics. Particularly suitable for scenarios involving continuous target variables, linear regression models the relationship between the predictors and the target using a linear equation. Despite its simplicity, linear regression offers valuable insights into the directional impact of predictors on the target variable, providing interpretable coefficients for each predictor. When appropriately applied, linear regression can serve as a reliable tool for imputing missing values, especially when the underlying relationship between variables is linear and well-defined.

Stacking, a sophisticated ensemble technique, holds promise as an effective modeling method in the context of predictive analytics. Stacking involves training multiple diverse models on the available dataset and using the predictions of these models as features for a meta-model, which then produces the final predictions. By leveraging the collective predictive power of diverse base models, stacking often outperforms individual models by capturing complementary information from different models, thereby reducing bias and variance in predictions. The versatility of stacking makes it a valuable tool for improving model performance across various analytical tasks.

Incorporating these diverse modeling methods into the analytical pipeline enriches the predictive capabilities, ultimately contributing to improved accuracy and reliability in handling complex datasets.

Given the multifaceted nature of Chicago's crime landscape and the intricate challenges associated with missing data imputation, we hypothesized that the integration of ensemble techniques, such as RandomForest and GradientBoosting, alongside advanced imputation tools like Multiple Imputation by Chained Equations (MICE), would lead to a significant improvement in imputation accuracy for Chicago arrest records. Furthermore, we anticipated that the utilization of Extreme Gradient Boosting (XGBoost) as a foundational model, coupled with ensemble methods, will result in enhanced predictive performance and robustness compared to singular imputation approaches. Through this research, the aim is to demonstrate that the innovative fusion of these methodologies will yield more reliable and precise imputation outcomes, thereby empowering stakeholders with actionable insights and bolstering decision-making capabilities in crime data analysis.

3 Methods

Our study centered on a dataset comprising crime data sourced from Illinois State, accessed via data.gov, serving as the foundational element for our imputation methodologies. We embarked on an exhaustive exploration intertwining advanced feature engineering with the predictive prowess of Random Forest and XGBoost technologies. Our overarching objective was to examine the intricate interplay between these methodologies and their collective impact on enhancing the precision of imputation, with the ambition of setting novel standards in data restoration accuracy.

To prepare the data, the dataset was initially loaded and inspected for basic understanding and to identify missing values. Boolean columns such as “Arrest” and “Domestic” were converted to integers to facilitate the analysis. Essential columns like “Beat”, “District”, “Ward”, “Community Area”, and “FBI Code” were converted to string type to maintain consistency and handle categorical data efficiently. This step ensured that the categorical variables were correctly processed during the imputation and modeling phases.

A critical step involved converting the “Date” column to datetime format, extracting day, month, year, and time components. The time was further converted to military time and categorized into parts of the day (e.g., Early Morning, Noon, Afternoon, Evening, Night) to capture temporal patterns in crime incidents. This temporal categorization allowed us to better understand the distribution of crimes throughout the day and identify potential trends.

Recognizing the importance of evaluating imputation methods under different conditions, we created three separate datasets with varying levels of missing data specifically for the "FBI Code" feature. These datasets were constructed with 10%, 25%, and 50% of the "FBI Code" data missing completely at random. This approach allowed us to systematically assess the impact of missing data rates on the performance of our imputation techniques and to determine whether the missing rate is a significant factor in imputation accuracy.

In addition, advanced feature engineering practices were applied, such as transforming categorical data and creating new time-based features. This included handling missing values using Random Forest and SimpleImputer techniques and ensuring the dataset was ready for model training. Feature engineering was crucial in enhancing the model’s ability to learn from the data and improve prediction accuracy.

We established a foundational imputation baseline model utilizing XGBoost. This model served as a benchmark to understand the initial performance without any optimizations. It provided a starting point for comparing the effectiveness of more complex techniques. Building on the baseline model, we integrated ensemble methods like Random Forest and Gradient Boosting. The ensemble approach aimed to combine the strengths of multiple models, mitigating overfitting and enhancing predictive accuracy. By leveraging the diverse strengths of each model, we aimed to create a more robust imputation framework.

Our evaluation was guided by the exclusive use of the accuracy score as our central metric due to the data being balanced. This metric simplified the evaluation process by providing a clear percentage of correct predictions, making it easy to understand and communicate the effectiveness of our approaches. Consistent use of this metric allowed for easy comparison between different models and approaches.

GridSearchCV was employed to systematically explore a wide range of hyperparameter combinations, enabling the identification of the most effective settings for our predictive model. This approach aimed to pinpoint the most relevant features and fine-tune our model parameters, thereby improving our capability to accurately predict the target values. By systematically testing various parameter combinations, we aimed to achieve the best possible performance from our models.

The accuracy metric was utilized with the balanced data by providing a clear percentage of correct predictions, making it easy to understand and communicate the effectiveness of our approaches. By focusing solely on the accuracy score, we streamlined our analysis, allowing for direct comparisons of model performance without the need for adjustments across different datasets. This approach emphasized the overall success rate of our models in making accurate predictions.

In tackling the identified challenges, our approach was shaped by feature engineering practices, such as one-hot-encoding, and the application of efficient modeling techniques. Leveraging advanced imputation tools such as XGBoost, an optimized distributed gradient boosting library, in tandem with an emphasis on ensemble methods, such as XGBoost and Random Forest, will serve as linchpins in navigating the complexities inherent in missing data. The integration of modeling methodologies such as GBM and enhanced Random Forest and Decision Trees, augmented by robust evaluation metrics including the Normalized Root Mean Square Error (NRMSE) and Proportion of Falsely Classified (PFC), facilitated a meticulous assessment of our strategies. This comprehensive approach aspires not only to address extant challenges but also to push the frontiers of current imputation practices, ensuring that our methodology stands at the vanguard of innovation while remaining deeply grounded in scholarly research.

4 Results

This section presents the results of our analysis of various imputation techniques applied to the Chicago crime dataset. We assess the impact of these methods on model accuracy using metrics such as Normalized Root Mean Square Error (NRMSE) and Proportion of Falsely Classified (PFC). By examining datasets with different levels of missing data, we demonstrate the effectiveness of ensemble methods and advanced imputation strategies in improving data quality and model performance. The results highlight the practical benefits of our approach for crime data analysis.

	Feature	Importance
0	FBI Code_18	0.071088
1	Primary Type_NARCOTICS	0.061936
2	IUCR_1811	0.028066
3	Latitude	0.022966
4	Year	0.022432
...	...	...
95	District_4.0	0.001497
96	IUCR_0910	0.001492
97	District_8.0	0.001478
98	Ward_37.0	0.001470
99	Ward_42.0	0.001464

100 rows x 2 columns

Fig. 1. Identifying features with the most influence on the target feature "Arrest"

Using a feature importance model, we identified the FBI Code as the most influential factor in determining arrests as shown above in Figure 1. To create a baseline model, we will impute the most important feature, the FBI Code, with new values. These imputed values will be integrated into the dataset alongside other features. Subsequently, we will predict the accuracy of arrests based on our new imputed values.

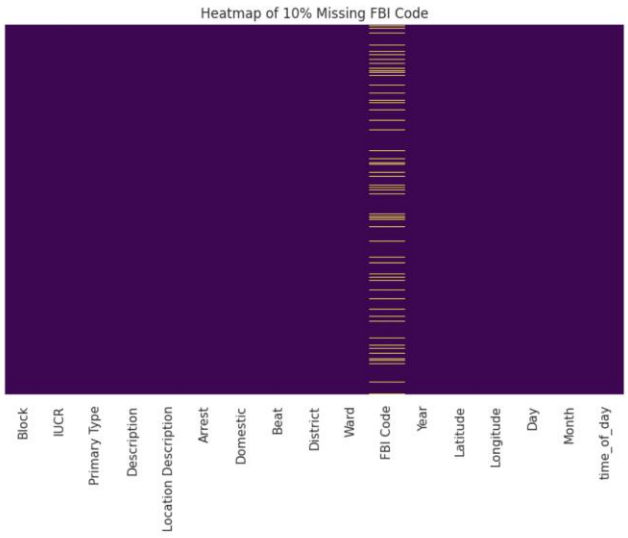

Fig. 2. Representation of FBI Code with 10% data missing at random

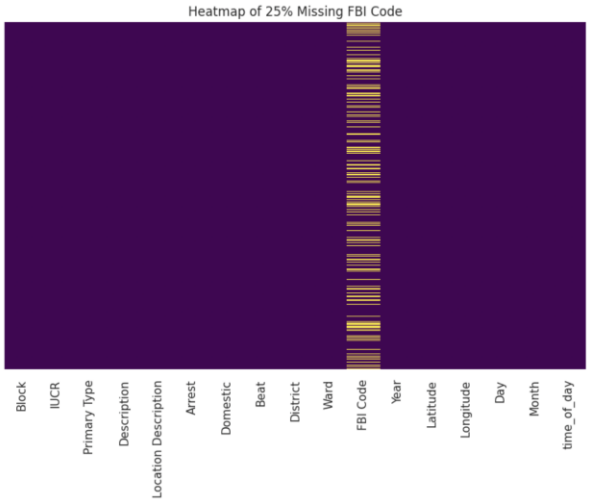

Fig. 3. Representation of FBI Code with 25% data missing at random

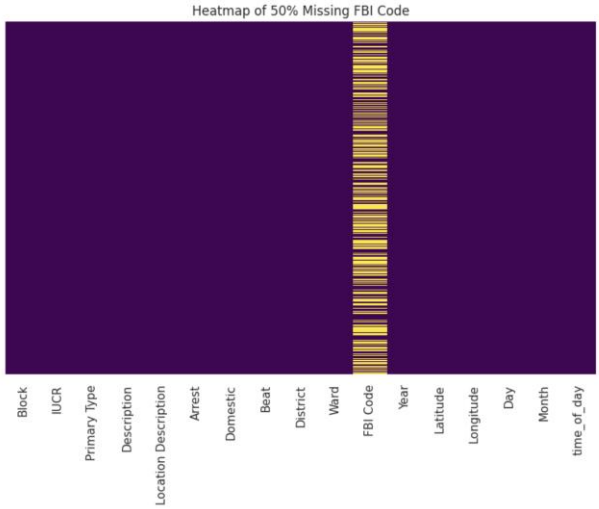

Fig. 4. Representation of FBI Code with 50% data missing at random

As previously mentioned, three different datasets were created, each with varying rates of missing data for the FBI Code: 10%, 25%, and 50%. This approach allowed for testing the effectiveness of the imputation model across different levels of missing data. The objective was to determine if the missing data was randomly distributed and to assess how well the imputation methods could handle these scenarios. In addition, the objective was also to determine if the missing rate had an impact on the methods described. As illustrated in Figures 2, 3, and 4, the datasets exhibited missing values completely at random as the percentage of missing data increased.

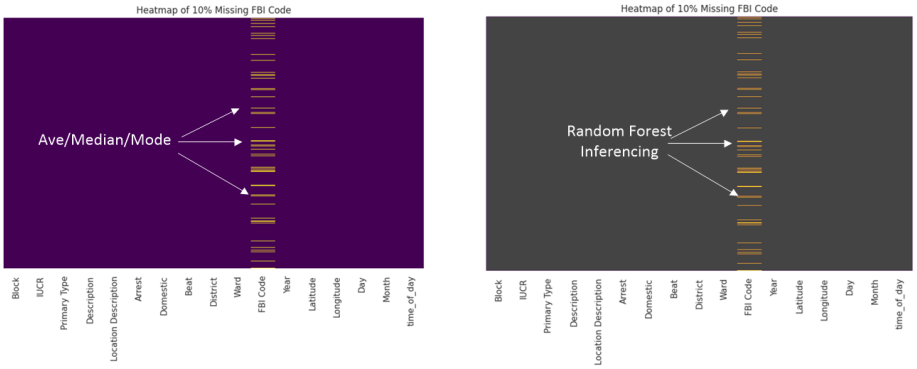

Fig. 5. Visual representation of the two imputation tools utilized. Simple imputer on the left and random forest imputer on the right.

Figure 5 provides a visual comparison of different imputation strategies applied to the FBI Code. The chart on the left employs the SimpleImputer, while the chart on the right utilizes the Random Forest Imputer. It is important to note that the Random Forest Imputer is distinct from the baseline random forest model used to predict arrests after imputing the FBI Code. Although this visualization specifically shows 10% missing values, the same imputation strategies were applied to datasets with 10%, 25%, and 50% missing values.

Model	10% Missing	25% Missing	50% Missing
Random Forest	88.3%	88.4%	88.4%
Decision Tree	85.6%	85.8%	85.6%
Gradient Boost	87.9%	87.9%	88.2%
Ensemble (RF+DT+GBM)	87.8%	87.8%	88.2%
Ensemble (RF+GBM)	88.1%	88.2%	88.6%

Fig. 6. Accuracy results of the feature "Arrest" with the substituted imputation feature "FBI Code"

With the three datasets featuring varying missing rates for the FBI Code—10%, 25%, and 50%—several singular models and a few ensemble models were run on each dataset. The results from the baseline models using simple imputation are as follows and represented in Figure 6: the Random Forest model consistently achieved an accuracy of approximately 88.3% to 88.4% across different missing rates. Decision Tree models displayed slightly lower accuracy, ranging from 85.6% to 85.8%. Gradient Boosting models had accuracies between 87.9% and 88.2%. The ensemble models, which combined Random Forest, Decision Trees, and Gradient Boosting, showed similar performance. Overall, the models performed similarly in terms of percentage, with no significant difference in accuracy as the missing data percentage increased. However, in real-world scenarios, even a 1% increase in accuracy can be highly beneficial and cost-saving for a company, highlighting the value of the imputation approach. Notably, the Random Forest model produced the best results, with a 0.5% to 1.5% increase in accuracy compared to other models, making it the preferred modeling approach for this crime dataset.

Random Forest Imputation Model	Results (Accuracy)
For 10% Missing	99.0%
For 25% Missing	98.5%
For 50% Missing	98.0%

Fig. 7. Accuracy results from the baseline simple imputer of "FBI Code"

The imputation model built using Random Forest demonstrated high accuracy in predicting the missing FBI codes. For datasets with 10% missing data, the model achieved an accuracy of 99.0%. With 25% missing data, the accuracy was 98.5%, and for 50% missing data, it was 98.0%. These high accuracy rates as shown in Figure 7 indicate the robustness of the imputation model, even as the proportion of missing data increased.

Model	SimpleImputer 10%	RF Imputer 10%	SimpleImputer 25%	RF Imputer 25%	SimpleImputer 50%	RF Imputer 50%
Random Forest	88.3%	88.4%	88.4%	88.5%	88.4%	88.7%
Decision Tree	85.6%	85%	85.8%	85.8%	85.6%	85.7%
Gradient Boost	87.9%	87.9%	87.9%	87.9%	88.2%	88.5%
Ensemble (RF+DT+GBM)	87.8%	87.7%	87.8%	87.9%	88.2%	87.7%
Ensemble (RF+GBM)	88.1%	88.1%	88.2%	88.4%	88.6%	88.5%

Fig. 8. Side by side comparison accuracy results of all modeling approaches with both simple imputer and random forest imputer.

The results in Figure 8 from both the Random Forest Imputer and the SimpleImputer are presented in the table. The Random Forest model consistently demonstrated high accuracy, with values ranging from 88.3% to 88.7% across different imputation methods and missing data percentages. The Decision Tree model exhibited slightly lower accuracy, around 85.6% to 85.8%, while the Gradient Boosting model achieved accuracies between 87.9% and 88.5%. The ensemble models, combining Random Forest, Decision Trees, and Gradient Boosting, showed similar performance, with accuracies ranging from 87.7% to 88.2%. Notably, the Random Forest Imputer, used for imputing missing data, proved effective in maintaining model performance, with the Random Forest model achieving the highest accuracy. Whether used alone or as part of an ensemble, the Random Forest model consistently produced the highest accuracy, demonstrating its robustness and effectiveness in handling varying levels of missing data. These results underscore the superiority of the Random Forest Imputer in maintaining model performance even with significant missing data.

5 Discussion

The results suggest that the research is on a promising path towards improving data imputation methods, with potential implications for various applications, especially those requiring high-quality data for accurate modeling, such as in the analysis of real-world crime records. The improved accuracy underscores the potential for enhanced decision-making capabilities based on more reliable data.

The significance of the research results lies in their potential to substantially enhance the quality and reliability of imputed data using novel ensembling techniques. This improvement in data quality is crucial across various fields, particularly in industries where decisions and outcomes are heavily reliant on accurate and complete datasets. For example, in healthcare, better imputation methods can lead to more accurate patient diagnoses and treatment plans, while in automotive safety, they can improve the analysis of car crash records, potentially leading to safer vehicle designs and road use policies. Moreover, the research provides a benchmark against which future imputation methods can be evaluated, fostering further innovation in the field. Data scientists and analysts are encouraged to integrate these new ensemble imputation techniques into their existing data processing workflows, thereby enhancing the integrity and reliability of their analyses. This, in turn, can lead to more informed decision-making and improved outcomes in various sectors. The broad applicability of these techniques highlights their transformative potential across multiple domains.

The innovative application of ensembling techniques to data imputation stands out as a novel approach, diverging from traditional methods and underscoring the significance of even modest improvements in data quality across various industries. This flexible strategy, which involves exploring and potentially combining multiple imputation methods, highlights an adaptive and innovative approach to enhancing data reliability. The anticipation that such improvements could significantly impact real-world outcomes, particularly in sectors like automotive, is both intriguing and unexpected. This research underscores the critical role of data quality in decision-making processes and the potential for innovative methodologies to drive significant advancements in diverse fields. The potential for significant practical impacts makes this research particularly noteworthy.

During the analysis, several challenges might be encountered, notably the complexity of the imputation model, which could present difficulties in its definition and potentially demand significant computational resources. Time constraints might also pose a challenge, especially in labor-intensive scenarios such as resampling data or when dealing with many variables. Additionally, the risk of overfitting could emerge as a concern, particularly with tools like MICE (Multiple Imputation by Chained Equations), which might be sensitive to model specifications. This issue could be compounded by convergence difficulties and instability in cases of multicollinearity, where variables are highly correlated, possibly complicating the imputation process further. Addressing these limitations will be crucial for the successful application of the proposed methods.

In advancing imputation methods, ethical considerations are paramount, particularly in ensuring transparency and preventing bias. The integrity of imputed data must be maintained to avoid misleading outcomes, especially when dealing with sensitive information. Ethical guidelines are essential to guide the responsible use of these techniques, safeguarding data privacy and maintaining trust in the analytical results derived from such methods. Specific guidelines to consider include ensuring that all personally identifiable information (PII) is anonymized or encrypted to protect

individuals' identities, complying with regulations such as GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act) to maintain data privacy, and implementing techniques to detect and mitigate biases in the data and the imputation process (Schubert et al, 2004). This includes checking for disparate impact on different demographic groups and ensuring that the imputed data does not reinforce existing biases. Maintaining clear documentation of the imputation methods used, including the assumptions made and the limitations of the techniques, is also crucial for transparency and accountability. This allows for reproducibility and accountability in the analytical process. Additionally, obtaining informed consent when using data from individuals is essential, ensuring that individuals are aware of how their data will be used, including imputation processes. Regular audits of the imputation processes and the models used should be conducted to ensure they adhere to ethical standards and perform as expected (Atlan, 2023). These audits should be documented and made available to relevant stakeholders. Evaluating the potential impact of imputed data on decision-making processes involves assessing the risks and benefits of using imputed data and ensuring that the outcomes are equitable and just (Atlan, 2023). By adhering to these guidelines, practitioners can safeguard data privacy, ensure the ethical use of imputation techniques, and maintain trust in the analytical results derived from such methods.

The field of data imputation is still in the early stages of development, offering ample opportunities for further exploration and refinement. As the discipline evolves, there is significant potential for researchers to delve deeper into advanced and more accurate imputation methods. The current state of research merely scratches the surface of what is possible, suggesting that future studies could uncover more nuanced and effective techniques. By employing a variety of innovative approaches and applications, there is hope that forthcoming research will mark a significant advancement in the realm of data imputation, pushing the boundaries of current methodologies and contributing to the enhancement of data quality across multiple domains. Future research will be vital in realizing the full potential of these innovative techniques.

6 Conclusion

This study showcased a novel method to address the longstanding issue of incomplete data by performing imputation using and combining ensemble methods along with extreme gradient boosting. When applied in the unique setting of Chicago crime data, the model's improved performance sets a benchmark for data imputation accuracy and brings new possibilities for this technique. The study's combination of analysis of existing methods and the extensive use of advanced techniques provides a compelling reference to overcome the challenges posed by missing datasets. This body of knowledge is not only useful for data science in general but also contributes to the improvement of integrity and robustness in data-based solutions and decision-making. By offering a robust and adaptable framework, this research paves the way for future advancements in data imputation.

References

1. Allison, P. (2001). Missing Data. https://statisticalhorizons.com/wp-content/uploads/2012/01/Milsap-Allison.pdf
2. Berrevoets, J., Imrie, F., Kyono, T., Jordon, J., & van der Schaar, M. (2023, February 24). To Impute or not to Impute? Missing Data in Treatment Effect Estimation. arxiv.org. https://arxiv.org/pdf/2202.02096.pdf
3. Chen, T., & Guestrin, C. (2016, June 16). XGBoost: A Scalable Tree Boosting System. arxiv.org. https://arxiv.org/pdf/1603.02754.pdf
4. Data Ethics Unveiled: Principles & Frameworks Explored. Atlan. (2023, November 28). https://atlan.com/data-ethics-101/#data-ethics--related-reads
5. Delgado-Panadero, Á., o Benítez-Andrades, J. A., & García-Ordás, M. T. (2023, July 5). A generalized decision tree ensemble based on the NeuralNetworks architecture: Distributed Gradient Boosting Forest (DGBF). www.arxiv.org. https://arxiv.org/ftp/arxiv/papers/2402/2402.03386.pdf
6. FBI. (2017, May 30). Fighting violent crime in Chicago. FBI. https://www.fbi.gov/video-repository/chicago_041817.mp4/view
7. Gurtskaia, K., Schwerter, J., & Doebler, P. (2024, January 26). Adapting tree-based multiple imputation methods for multi-level data? A simulation study. browse.arxiv.org. https://browse.arxiv.org/pdf/2401.14161.pdf
8. Guthmann, A. (2024, March 11). As Homicides Drop Nationwide, Chicago Lags Behind Other Major Cities. Why?. WTTW. https://news.wttw.com/2024/03/11/homicides-drop-nationwide-chicago-lags-behind-other-major-cities-why
9. Hsu, C.-H., & Yu, M. (2017, October 12). Cox regression analysis with missing covariates via multiple imputation. arXiv.org. https://arxiv.org/abs/1710.04721
10. Kavelaars, X. M., Van Buuren, S., & Van Ginkel, J. R. (2019, April 8). Multiple imputation in data that grow over time: A comparison of three strategies. https://arxiv.org/pdf/1904.04185.pdf
11. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neurips.cc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
12. Manski, C. F. (2022, May 2). INFERENCE WITH IMPUTED DATA: The Allure of Making Stuff Up. arXiv.org. https://arxiv.org/ftp/arxiv/papers/2205/2205.07388.pdf
13. Nile, T., Qin, G., & Sun, J. (2022, May 20). Truncated tensor Schatten p-norm based approach for spatiotemporal traffic data imputation with complicated missing patterns. https://arxiv.org/pdf/2205.09390.pdf
14. Nguyen, T., Vo, T. L., Halvorsen, P., & Riegler, M. A. (2024, January 28). Imputation using training labels and classification via label imputation. browse.arxiv.org. https://browse.arxiv.org/pdf/2311.16877.pdf
15. Schubert, K. D., & Barrett, D. (2024). Human privacy in virtual and Physical Worlds: Multidisciplinary perspectives. Springer Nature Switzerland Palgrave Macmillan.
16. Stempfle, L., Panahi, A., & Johansson, F. (2023, November 24). Sharing Pattern Submodels for Prediction with Missing Values. https://arxiv.org/pdf/2206.11161v3.pdf
17. Thurow, M., Dumpert, F., & Pauly, M. (2021, January 19). Goodness (of fit) of Imputation Accuracy: The GoodImpact Analysis. arxiv.org. https://arxiv.org/pdf/2101.07532.pdf

18. Vallas, P. (2023, July 27). Vallas: Here's where Chicago Board of Education should put its focus. Illinois Policy. https://www.illinoispolicy.org/vallas-heres-where-chicago-board-of-education-should-put-its-focus/
19. Van der Laan, M., Polley, E., & Hubbard, A. (2007, September 16). Super Learner (Abstract). https://pubmed.ncbi.nlm.nih.gov/17910531/
20. Younus, S., Rönnstrand, L., & Kazi, J. U. (2018). Xputer: Bridging Data Gaps with NMF, XGBoost, and a Streamlined GUI Experience. https://arxiv.org/ftp/arxiv/papers/2311/2311.03747.pdf