

A new data imputation technique for efficient used car price forecasting

Charlène Béatrice Bridge-Nduwimana¹, Aziza El Ouazizi^{1,2}, Majid Benyakhlef²

¹Laboratory for Artificial Intelligence, Data Science and Emerging Systems, Fes National School of Applied Sciences, Sidi Mohamed Ben Abdellah University, Fes, Morocco

²Laboratory of Engineering Sciences, Polydisciplinary Faculty of Taza, Sidi Mohamed Ben Abdellah University, Morocco

Article Info

Article history:

Received Jul 6, 2024

Revised Nov 9, 2024

Accepted Nov 20, 2024

Keywords:

Feature engineering

Imputation

Missing values

Preprocessing

Regression

Used car forecasting

ABSTRACT

This research presents an innovative methodology for addressing missing data challenges, specifically applied to predicting the resale value of used vehicles. The study integrates a tailored feature selection algorithm with a sophisticated imputation strategy utilizing the *HistGradientBoostingRegressor* to enhance efficiency and accuracy while maintaining data fidelity. The approach effectively resolves data preprocessing and missing value imputation issues in complex datasets. A comprehensive flowchart delineates the process from initial data acquisition and integration to ultimate preprocessing steps, encompassing feature engineering, data partitioning, model training, and imputation procedures. The results demonstrate the superiority of the *HistGradientBoostingRegressor* for imputation over conventional methods, with boosted models extreme gradient boosting (XGBoost) regressor and gradient boosting regressor exhibiting exceptional performance in price forecasting. While the study's potential limitations include generalizability across diverse datasets, its applications include enhancing pricing models in the automotive sector and improving data quality in large-scale market analyses.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Charlène Béatrice Bridge-Nduwimana

Laboratory for Artificial Intelligence, Data Science and Emerging Systems, Fes National School of Applied Sciences, Sidi Mohamed Ben Abdellah University

Fes, Morocco

Email: charlenebeatrice.bridgenduwimana@usmba.ac.ma

1. INTRODUCTION

Handling massive datasets with missing values is a prevalent challenge in modern data analysis, necessitating innovative solutions. Missing values often result from the exponential growth of data sources and the inherent incompleteness of data collection processes. Advanced technologies and analytical methods, including machine learning and statistical techniques, are crucial for managing large-scale data efficiently and imputing missing values effectively. The complexity and volume of data require robust and scalable solutions to ensure accurate analysis and derive meaningful insights. Recently, predicting the resale value of used vehicles (PRVUV) has gained importance. Used car sales represent a significant sector, yet forecasting demand remains challenging due to various factors, such as unique characteristics, limited data availability, and dynamic market conditions. Accurate forecasting methods are essential for optimizing sales and inventory management.

Current research on imputation methods for missing data highlights the evolving landscape of data preprocessing techniques. Modern approaches focus on preserving statistical relationships and accounting for uncertainty within the data. Simply discarding incomplete cases can lead to a significant loss of valuable

information. Therefore, employing multivariable imputation techniques, which produce multiple imputations based on other observed attributes, is advisable. Studies [1], [2] demonstrate that advanced methods, such as multiple imputation by chained equations (MICE) and machine learning-based approaches like random forest (RF) imputation and k-nearest neighbors (KNN) imputation, often outperform traditional methods in managing complex missing data patterns. These techniques have shown promise in maintaining statistical power and reducing bias in subsequent analyses, particularly in large-scale datasets with mixed variable types. Furthermore, meticulous feature selection [3] is critical for price prediction in regression models using machine learning algorithms. To underscore the importance of feature selection, methods like semi-logarithmic hedonic regression [4] have been used in the literature to identify characteristics with positive effects, including diesel engines, specific colors (black and grey), automatic transmissions, country and year of manufacture, sunroofs, and engine cylinder specifications.

Research has proposed various solutions, such as multi-stage systems that offer functionalities like website filtering, training, data preparation, prediction, and vehicle state adjustments to address irregularities [5]. These systems utilize rule-based integration and encompass evaluation frameworks [6], market value estimation for used vehicles [7], fair price forecasting models [8], and video-based car model detection systems [9]. Additionally, diverse applications demonstrate the extensive use of machine learning in the automotive industry. In studies related to PRVUV, it has been observed that forecasting can be effectively conducted using traditional regression models, such as artificial neural networks, boosted models, linear regression, and regression trees [10]–[12]. However, decision trees and naïve Bayes models are generally not suitable for continuous-valued data when the number of observations is small.

The challenge of missing data can significantly impact the accuracy and efficacy of analytical tasks and statistical analyses. This issue is addressed in section 2, where we introduce a feature selection method as a preprocessing step and discuss imputation techniques for regression analysis on a dataset of used cars in the United States. Our proposed framework combines these strategies to effectively handle missing data. A key contribution, detailed in this section, is the adaptation of a specific structure to manage missing values in a machine learning project. We propose a two-step process: first, utilizing an optimal predictor for observed data without missing values to sequentially predict variables with missing values and reintegrate them into the initial dataset; second, applying a combination of imputation methods depending on the variable type (numerical or categorical). This pre-learning imputation is a notable aspect of our approach with significant practical implications, which we will discuss further. Section 3 explores the performance results of the various methods employed in section 2, and section 4 concludes our study, highlighting the effectiveness of the proposed imputation technique and providing future directions for further enhancing the imputation process.

2. METHODS AND MATERIALS

2.1. Data preprocessing

Understanding the dataset is crucial for any data science project. This process involves defining the data source, exploring its attributes, and analyzing feature relationships to identify underlying patterns or correlations. Statistical and machine learning algorithms [13], [14] can provide insights to guide the preprocessing and modelling stages. It is essential to perform certain preprocessing actions before imputation to address specific limitations [15]–[17], such as high sensitivity to outliers, selection of conditioning variables affected by imputation methods, and transformation of ordinal or categorical values. The dataset initially consists of 3,000,040 entries and 66 features, sourced from Kaggle dataset (<https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset>). The data was collected using a custom web crawler that extracted information from the Cargurus inventory in September 2020, making it suitable for experimental purposes.

To optimize our regression model for vehicle price prediction, we implemented a feature selection method with multiple functionalities aimed at identifying and removing redundant or irrelevant features. This approach addresses the critical question of which vehicle attributes should be included in the model:

- a. Identifying columns with a high percentage of missing data (above a 20% threshold),
- b. Detecting columns with a single unique value finding collinear variables with a high correlation coefficient (above a 90% threshold),
- c. Combining these conditions to identify features for removal,
- d. Removing the identified features,
- e. Visualizing the data using histograms for missing values and unique values, and heatmaps for further insights.

Subsequently, we applied consistent formatting to features, used standardization or normalization techniques for numerical variables, and applied appropriate encoding methods for categorical features to ensure data consistency and compatibility for model training.

2.2. Imputation prior to analysis

Most statistical models and machine learning algorithms are not designed to handle incomplete data. There are three mechanisms of missing data [18]: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR indicates that missingness is independent of both observed and unobserved data. In contrast, MAR and MNAR imply that missingness is related to the observed data or the missing values themselves. Categorizing missing data is challenging since missing values often relate to non-missing variables. It is generally advisable to treat missing data as MAR, positioned between these mechanisms. The selection of imputation methods depends on the mechanisms and patterns of missing data. No single method is suitable for all scenarios. Imputation techniques are categorized as single (e.g., mean/mode/constant, regression, and hot-deck) or multiple. Single imputation replaces each missing value once, as in studies [17], [19], [20], which include methods like iterative imputer, KNN imputer, K-means clustering, mean imputation, and decision trees (CART). Multiple imputation, on the other hand, creates several datasets with different imputed values that are later combined. An example of multiple imputation can be found in [21], where the class center missing value imputation (CCMVI) method is enhanced and merged with other approaches, such as imputing based on the nearest class center and using the mean of class centers to address missing values in the test dataset.

For our study, we explored the multiple imputation approach. The notation of each experiment is in the form “(numerical imputation_categorical imputation)” to specify the imputation method by data type. For all categorical values, we applied *SimpleImputer()* where values are replaced by their mode.

a. *SimpleImputer()_SimpleImputer() (Simp_Simp)*:

We replaced both numerical and categorical missing data using *SimpleImputer()* [20]. A straightforward and widely used method. Particularly for categorical variables where the most frequent value (mode) is substituted.

b. *IterativeImputer(LinearRegression())_SimpleImputer() (Iter_Simp)*:

IterativeImputer(LinearRegression()) [19] is a flexible tool offering various estimation strategies. It performs iterative imputation for numerical values, refining values until convergence or a specified maximum number of iterations, using *LinearRegression()* as the estimator. Categorical missing values are imputed using *SimpleImputer()*.

c. *KNNImputer()_SimpleImputer() (KNN_Simp)*:

KNNImputer() [19] identifies the K nearest neighbors (we use $k=5$) by computing the distance between incomplete and complete data points, typically using Euclidean distance. It estimates the numerical missing values based on these K neighbors' values, providing an effective solution.

d. *HistGradientBoostingRegressor()_SimpleImputer() (Ours_Simp)*:

This regression imputation method uses complete data to formulate regression equations, which are then employed to predict and fill in missing data values. We selected the *HistGradientBoostingRegressor()* regression method, a rarely explored solution for data imputation in the literature, as cited in [22]–[24].

“Ours” as shown in Figure 1 is explained below:

a. Dataset preparation:

- Feature engineering: Apply the feature selection method to remove 26 detected features
- Data organization: Standardize numerical variables and apply *LabelEncoder()* for categorical variables
- *Dataframe* split: create one dataset with missing values and another without missing values.

b. Imputation process: For each missing feature (“maximum_seating”, “engine_displacement”, “horsepower”, “mileage”, “seller_rating”):

- Separate the target variable from features in the non-missing dataset
- Train 10 *HistGradientBoostingRegressor* models with specified parameters ($learning_rate = 0.1$, $max_iter = 100$, $max_depth = 15$, $min_samples_leaf = 30$, $l2_regularization = 0.1$, $max_bins = 128$, $early_stopping = True$, $validation_fraction = 0.15$, $n_iter_no_change = 10$, and $scoring = 'neg_mean_squared_error'$)
- Predict missing values using trained models on the dataset with missing values (excluding one predicted feature)
- Average the predictions from all models
- Fill in missing values in the original *dataframe*
- Concatenate the filled *dataframe* with the non-missing *dataframe*
- Sort by index to restore the original order

c. Final preprocessing: Based on a literature review, we removed 14 additional features, resulting in a total of 40 features removed.

2.3. Learning models and performance metrics

The proposed model is designed to address complex, non-linear relationships that basic regression methods often fail to capture, while ensuring a reasonable degree of stability. Our approach utilizes a range of advanced prediction techniques, including ensemble methods, tree-based algorithms, and boosting strategies. To evaluate the effectiveness and robustness of these models when applied to our dataset, we perform a comparative analysis of six models, alongside a linear-based ridge regression model serving as a baseline.

- a. Extra trees regressor (ExtraTrReg) [10]: An ensemble method that aggregates fully grown decision trees to reduce variance and bias. It is computationally efficient, suitable for large datasets, and demonstrates high accuracy.
- b. Bagging regressor (BaggReg) [25]: An ensemble method that reduces variance and is robust against overfitting, making it particularly effective for high-variance, low-bias models like decision trees.
- c. eXtreme gradient boosting regressor (XGBReg) [22], [23], [25]: A scalable and efficient implementation of gradient boosting, focusing on speed and performance. It offers regularization, parallel processing, internal handling of missing values, and customizable hyperparameters.
- d. Ridge regressor (Ridge) [22]: A linear regression model with L2 regularization that mitigates overfitting and is suitable for high-dimensional data. It helps address multicollinearity by adding a penalty to the size of the coefficients.
- e. Decision tree regressor (DecisionTr) [10], [25], [26]: A non-parametric model that splits the data into subsets based on input features, forming a tree where each leaf represents predicted values. It is easy to interpret and can capture non-linear relationships but is prone to overfitting.
- f. Gradient boosting regressor (GBReg) [25]–[27]: An additive model built in a forward stage-wise manner, optimizing for a differentiable loss function. It is highly accurate but can overfit if not properly regularized, making it suitable for regression problems requiring high predictive performance.
- g. Random forest regressor (RF) [25], [26]: A decision tree ensemble trained using the bagging method, where each tree is trained on a bootstrap sample of the data. It reduces overfitting, is robust to noise, and provides feature importance scores.

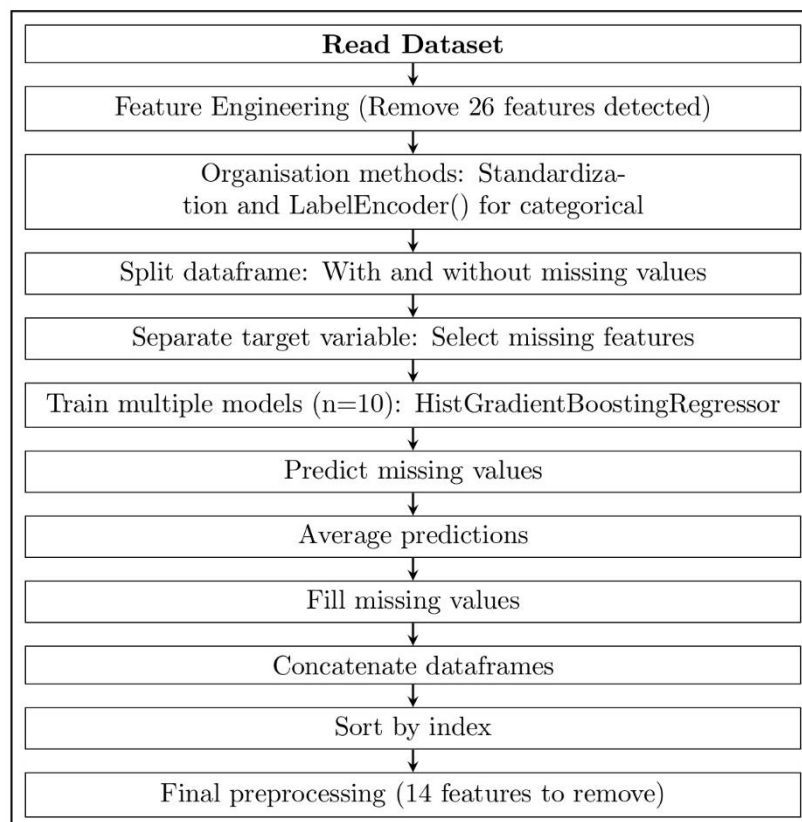


Figure 1. Flowchart of the proposed data imputation model

For performance evaluation [1], we use the R2-score regression metric as in (1), which is a measure of how well the model explains the variability of the response variable. An R2-score close to 100% indicates a high level of accuracy in predictions. However, this metric can be difficult to interpret on its own and should be complemented with other metrics [26] such as mean squared error (MSE) or mean absolute error (MAE), which are useful in regression tasks. In our study, we employed the root mean square error (RMSE) metric as in (2) to quantify the differences between the predicted and actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

The two fundamental equations for the metrics considered: (1) R2-score and (2) RMSE where n is the number of observations, y_i is the observed value, \bar{y} is the average value, and \hat{y}_i is the predicted value.

3. RESULTS AND DISCUSSION

3.1. Results

In our experiments, we utilized a dataset with missing values and conducted the analysis using Python. The original dataset comprised 3,000,040 rows and 66 columns, which, after feature engineering and data preparation, was reduced to 26 features. The prediction tasks were defined based on 25 input features and 1 outcome variable. To ensure consistency, the dataset was subsampled into five equal parts of 600,008 samples each. One sample was used to perform a train-test split (70/30) using the *train_test_split* function from Scikit-Learn. The results presented in Tables 1-2 and Figures 2(a)-2(b) and 3(a)-3(b) illustrate the performance of the various models in predicting used car prices, under different imputation techniques.

Table 1. R2Score in % (RMSE) for training set

Imputations methods	Regression learning machines results						
	ExtraTrReg	BaggReg	XGBReg	Ridge	DecisionTr	GBReg	RF
<i>Simp_Simp</i>	83.75 (0.0766)	91.01 (0.0570)	97.42 (0.0305)	70.95 (0.1017)	91.27 (0.0562)	92.99 (0.0503)	88.34 (0.0649)
<i>Iter_Simp</i>	83.77 (0.0766)	90.76 (0.0578)	97.47 (0.0303)	70.77 (0.1014)	91.60 (0.0551)	93.12 (0.0499)	88.71 (0.0639)
<i>KNN_Simp</i>	83.75 (0.0766)	90.98 (0.0571)	97.37 (0.0308)	71.10 (0.1014)	91.71 (0.0547)	93.37 (0.0490)	89.06 (0.0629)
<i>Ours_Simp</i>	84.13 (0.0757)	92.75 (0.0512)	97.91 (0.0275)	72.17 (0.0995)	91.89 (0.0541)	96.51 (0.0355)	90.44 (0.0588)

Table 2. R2Score in % (RMSE) for testing set

Imputations methods	Regression learning machines results						
	ExtraTrReg	BaggReg	XGBReg	Ridge	DecisionTr	GBReg	RF
<i>Simp_Simp</i>	83.29 (0.0758)	87.06 (0.0667)	89.34 (0.0606)	74.33 (0.0930)	84.88 (0.0721)	87.77 (0.0649)	84.00 (0.0742)
<i>Iter_Simp</i>	83.18 (0.0761)	86.54 (0.0680)	89.18 (0.0610)	74.14 (0.0928)	85.50 (0.0706)	88.97 (0.0616)	85.07 (0.0717)
<i>KNN_Simp</i>	83.05 (0.0764)	86.82 (0.0673)	90.24 (0.0580)	74.56 (0.0926)	84.71 (0.0725)	87.64 (0.0652)	84.16 (0.0738)
<i>Ours_Simp</i>	83.74 (0.0748)	88.18 (0.0638)	90.73 (0.0565)	75.64 (0.0906)	85.26 (0.0712)	90.68 (0.0566)	85.21 (0.0713)

3.2. Discussion

This study introduces a novel approach to managing missing data in large datasets, specifically focusing on the prediction of used car prices. The methodology integrates a custom feature selection technique with an advanced imputation technique using *HistGradientBoostingRegressor()*. By combining these elements, the research aims to enhance the efficiency and accuracy of handling missing data while preserving dataset integrity. The study's originality lies in exploring the unique capabilities of *HistGradientBoostingRegressor()* for directly managing missing values. A comparative analysis of regression models on complete data was conducted to evaluate the effectiveness of this approach.

From the performance metrics of tree-based models as shown in Tables 1 and 2, ExtraTrReg was the least effective, while BaggReg performed better than DecisionTr and RF. Among all the evaluated models, the boosted-based models achieved the highest performance scores. Concerning the existing imputation methods (*Simp_Simp*, *Iter_Simp*, and *KNN_Simp*), all models showed similar performance on both training and test sets. However, our proposed imputation method created a noticeable performance gap, particularly for the BaggReg and GBReg models. The Ridge regression model, considered as a baseline, proved to be the least effective, indicating that the task of predicting used car prices involves non-linearity, rendering linear

regression approaches less suitable despite preprocessing adjustments. Ultimately, three models (RF, BaggReg, and GBReg) stood out in terms of their performance as indicated by the R2-Score and RMSE metrics. These performance metrics were significantly improved with our proposed imputation method (*Ours_Simp*), as depicted in Figures 2(a)-2(b) and 3(a)-3(b).

The study’s innovative approach to handling missing data in car price prediction shows promise, though it may face limitations in its generalizability across diverse datasets. The effectiveness of the custom feature selection and the *HistGradientBoostingRegressor* based imputation may vary with different data distributions or missing data mechanisms (MCAR and MNAR). Further research is required to validate the robustness of this method across various domains.

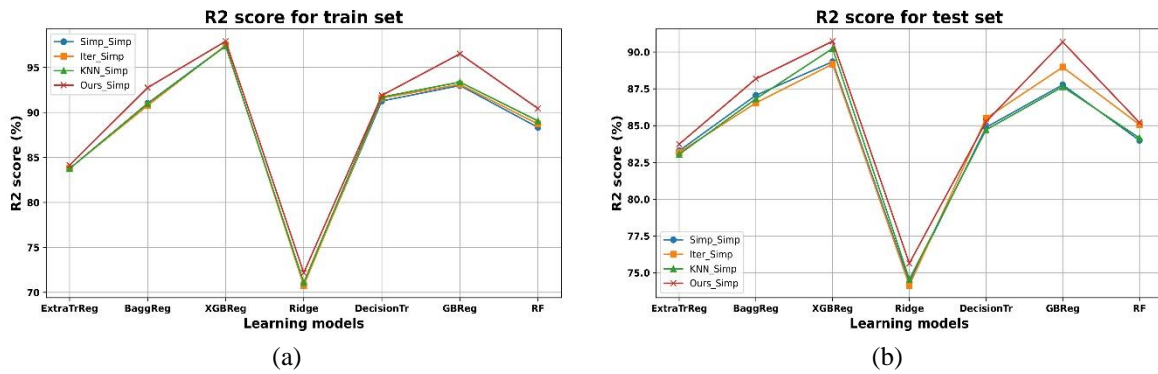


Figure 2. Analyzing predictive accuracy for price estimation by contrasting outcomes from imputation methods applied to missing data in both (a) training and (b) testing sets

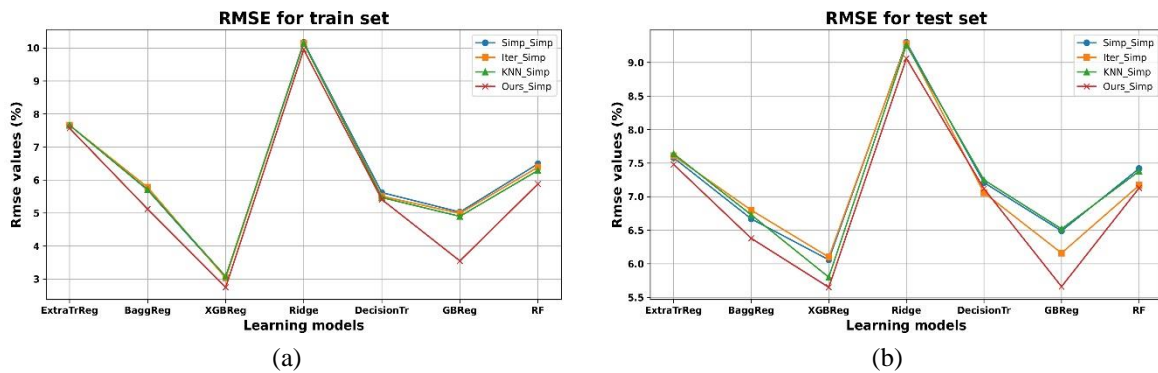


Figure 3. RMSE values illustrating the impact of various imputation techniques on both (a) the training dataset and (b) the testing dataset




4. CONCLUSION

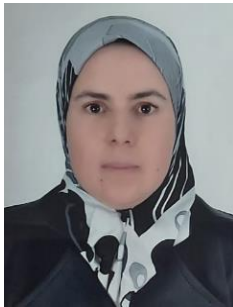
This study aimed to address the challenge of missing data in large datasets and improve the accuracy of used car price predictions. The effectiveness of the proposed imputation method is noted to depend on data distributions and missing data patterns. The results demonstrate that our imputation method enhances the performance of boosted-based models in terms of R2-Score and RMSE metrics, especially for the eXtreme Gradient Boosting and Gradient Boosting Regression models, compared to existing imputation methods. Practical applications of this study can be found in the automotive industry for optimizing used car price prediction and inventory management. Additionally, industries facing similar challenges with large datasets and missing data, such as healthcare, finance, and insurance, can benefit from accurate predictions and insights to improve decision-making. The study achieved comparable performance to existing literature. Future research will focus on improving the proposed imputation technique, optimizing hyperparameters, and exploring other machine learning approaches, including deep learning models. Further research directions include validating the robustness of the proposed imputation technique across varying data distributions and missing data mechanisms.




REFERENCES

- [1] J. H. Li *et al.*, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Medical Research Methodology*, vol. 24, no. 1, 2024, doi: 10.1186/s12874-024-02173-x.
- [2] M. N. Arefin and A. K. M. Masum, "A probabilistic approach for missing data imputation," *Complexity*, vol. 2024, 2024, doi: 10.1155/2024/4737963.
- [3] A. Matas and J. L. Raymond, "Hedonic prices for cars: an application to the Spanish car market, 1981-2005," *The Applied Economics of Transport*, pp. 99–116, 2014, doi: 10.4324/9781315872360-8.
- [4] C. Erdem and I. Şentürk, "A hedonic analysis of used car prices in Turkey," *International Journal of Economic Perspectives*, vol. 3, no. 2, pp. 141–149, 2009.
- [5] A. Ifthikar and K. Vidanage, "Valuation of used vehicles: a computational intelligence approach," in *Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS*, 2018, vol. 2018-May, pp. 7–10, doi: 10.1109/ISMS.2018.00011.
- [6] B. C. Zavitz, P. A. Russek, J. I. Puente, and N. J. Park, "Systems and methods for facilitating the purchase of one or more vehicles," AutoTrader Inc., Patent no. 10102555, Oct. 16, 2018.
- [7] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car price prediction using machine learning techniques," *TEM Journal*, vol. 8, no. 1, pp. 113–118, 2019, doi: 10.18421/TEM81-16.
- [8] C. Longani, S. P. Potharaju, and S. Deore, "Price prediction for pre-owned cars using ensemble machine learning techniques," *Advances in Parallel Computing*, vol. 39, pp. 178–187, 2021, doi: 10.3233/APC210194.
- [9] K. B. Chigateri, S. Suryavamshi, and S. Rajendra, "System for detecting car models based on machine learning," *Materials Today: Proceedings*, vol. 52, pp. 1697–1701, 2022, doi: 10.1016/j.matpr.2021.11.335.
- [10] C. L. Lasya, S. Pooja, S. Jeyashree, C. Ambhika, and G. Eswari, "Forecasting Pre-Owned Car Prices Using Machine Learning," in *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2023*, 2023, pp. 1–6, doi: 10.1109/ICSTSN57873.2023.10151632.
- [11] P. Srinivasan, R. O. Reddy, K. A. Sai, and J. Naidu, "Predictive analysis and application of various machine learning algorithms to forecast used car prices," in *International Conference on Sustainable Computing and Smart Systems, ICSCSS 2023 - Proceedings*, 2023, pp. 190–194, doi: 10.1109/ICSCSS57650.2023.10169183.
- [12] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buaya, and P. Boonpou, "Prediction of prices for used car by using regression models," in *Proceedings of 2018 5th International Conference on Business and Industrial Research: Smart Technology for Next Generation of Information, Engineering, Business and Social Science, ICBIR 2018*, 2018, pp. 115–119, doi: 10.1109/ICBIR.2018.8391177.
- [13] H. M. Safhi, B. Frikh, B. Hirchoua, B. Ouhbi, and I. Khalil, "Data intelligence in the context of big data: a survey," *Journal of Mobile Multimedia*, vol. 13, no. 1–2, pp. 1–27, 2017.
- [14] V. Viswanatha, A. C. Ramachandra, B. D. Parameshchari, H. V. Vachan, and S. S. Shetty, "Predicting the price of used cars using machine learning," in *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques, EASCT 2023*, 2023, pp. 1–6, doi: 10.1109/EASCT59475.2023.10393486.
- [15] M. G and N. G, "A new paradigm for development of data imputation approach for missing value estimation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, no. 6, pp. 3222–3228, 2016, doi: 10.11591/ijece.v6i6.pp3222-3228.
- [16] K. Trang, A. H. Nguyen, L. Tonthat, and B. Q. Vuong, "Improving RepVGG model with variational data imputation in COVID-19 classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 4, pp. 1278–1286, 2022, doi: 10.11591/ijai.v11.i4.pp1278-1286.
- [17] A. Salem, N. A. Emran, A. K. Muda, Z. Sahri, and A. Ali, "Missing values imputation in Arabic datasets using enhanced robust association rules," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 28, no. 2, pp. 1067–1075, 2022, doi: 10.11591/ijeecs.v28.i2.pp1067-1075.
- [18] G. Papageorgiou, S. W. Grant, J. J. M. Takkenberg, and M. M. Mokhles, "Statistical primer: how to deal with missing data in scientific research?," *Interactive Cardiovascular and Thoracic Surgery*, vol. 27, no. 2, pp. 153–158, 2018, doi: 10.1093/icvts/ivy102.
- [19] N. Mandal and T. Sarode, "A framework for cloud cover prediction using machine learning with data imputation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 1, pp. 600–607, 2024, doi: 10.11591/ijece.v14i1.pp600-607.
- [20] S. Lestari, Yulmaini, Aswin, S. Y. Ma'ruf, Sulyono, and R. R. N. Fikri, "Alleviating cold start and sparsity problems in the micro, small, and medium enterprises marketplace using clustering and imputation techniques," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 3, pp. 3220–3229, 2024, doi: 10.11591/ijece.v14i3.pp3220-3229.
- [21] Y. Hanyf and H. Silkan, "A method for missing values imputation of machine learning datasets," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, pp. 888–898, 2024, doi: 10.11591/ijai.v13.i1.pp888-898.
- [22] A. Perez-Lebel, G. Varoquaux, M. Le Morvan, J. Josse, and J. B. Poline, "Benchmarking missing-values approaches for predictive models on health databases," *GigaScience*, vol. 11, 2022, doi: 10.1093/gigascience/giac013.
- [23] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00516-9.
- [24] S. Li, "Estimating stock market prices with histogram-based gradient boosting regressor: a case study on Alphabet Inc.," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, pp. 532–543, 2024, doi: 10.14569/IJACSA.2024.0150553.
- [25] A. Ayaou, "Used car price prediction in Morocco using machine learning," Mohamed I University, 2022.
- [26] A. Kumar, "Machine learning based solution for asymmetric information in prediction of used car prices," *International Conference on Intelligent Vision and Computing*, pp. 409–420, 2023, doi: 10.1007/978-3-031-31164-2_34.
- [27] M. Kaur, S. Singh, and N. Aggarwal, "Missing traffic data imputation using a dual-stage error-corrected boosting regressor with uncertainty estimation," *Information Sciences*, vol. 586, pp. 344–373, 2022, doi: 10.1016/j.ins.2021.11.049.




BIOGRAPHIES OF AUTHORS

Charlène Béatrice Bridge-Nduwimana    received the M.S. degree in communication, telecommunication systems, and computer networks from Moulay Slimane University of Beni Mellal, Morocco, and a B.Sc. and Technology degree in engineering sciences from University Moulay Ismail, Faculty of Sciences and Technology of Errachidia. Currently, she is a Ph.D. student at the University of Sidi Mohamed Ben Abdellah in Fes, Faculty of Sciences and Technology. Data science, optimization, machine learning, and artificial intelligence are among her areas of interest in study. She can be contacted at email: charlenebeatrice.bridgenduwimana@usmba.ac.ma.



Aziza El Ouaazizi    holds her Ph.D. at Sidi Mohamed Ben Abdellah University in 2000. After working as a professor in Technical High School of Fes (2001), she is currently working as professor in the Informatics at Sidi Mohamed Ben Abdellah University, Fez. She is also a permanent member of Artificial Intelligence Data Sciences and Emergent Systems Laboratory and an associate member of Engineering Science Laboratory. Her research interests include machine and deep learning, artificial vision and image processing, pattern recognition, data analysis, evolutionary algorithms and their applications. She can be contacted at email: aziza.elouaazizi@usmba.ac.ma.



Majid Benyakhlef    received a Ph.D. degree by Sidi Mohamed Ben Abdellah University, Faculty of Science in Fes, Morocco. He works as an associate professor of informatics and automatic control at Polydisciplinary Faculty of Taza. He is also a permanent member of Engineering Science Laboratory. His current areas of interest in research are large-scale adaptive control, fuzzy control, and decentralized robust control, computer science in the fields of modelling, analysis and information processing. He can be contacted at email: majid.benyakhlef@usmba.ac.ma.