Project Report: Missing Values Imputation in Car Price Data:

By Deepak & Pankaj

Project Overview:

The objective of this project was to explore different techniques for handling missing values in a dataset and compare their effectiveness. We worked with a curated dataset on car prices, intentionally removing 20 values to create a dataset with missing data. The goal was to test the following imputation techniques:

1. SimpleImputer with mean, median, and most frequent strategies
2. Linear Regression
3. k-Nearest Neighbors (KNN)
4. Multivariate Imputation by Chained Equations (MICE)

Dataset Description:

The dataset contains information on various car models, their features, and their prices. The features included in this dataset are:

1. car_ID: Unique identifier for each car
2. symboling: Risk factor assigned by car insurance companies
3. CarName: Name of the car model
4. fueltype: Type of fuel used (e.g., gas, diesel)
5. aspiration: Type of engine aspiration (e.g., turbo, standard)
6. doornumber: Number of doors (two or four)
7. carbody: Type of car body (e.g., sedan, hatchback)
8. drivewheel: Type of drive wheel (e.g., FWD, RWD)
9. enginelocation: Location of the engine (e.g., front, rear)
10. wheelbase: Distance between front and rear wheels
11. carlength: Length of the car
12. carwidth: Width of the car
13. carheight: Height of the car
14. 1curbweight: Weight of the car without passengers or cargo
15. enginetype: Type of engine (e.g., OHV, DOHC)
16. cylindernumber: Number of cylinders
17. 1enginesize: Engine displacement
18. fuelsystem: Fuel system type (e.g., MPFI, 2bbl)
19. 1boreratio: Ratio of bore diameter to stroke
20. stroke: Length of piston stroke
21. compressionratio: Compression ratio of the engine
22. horsepower: Power output of the engine
23. peakrpm: RPM at peak horsepower
24. citympg: City fuel economy
25. highwaympg: Highway fuel economy
26. price: Price of the car

We removed 20 values from this dataset to simulate real-world scenarios where missing data is common. These missing values were randomly distributed across different features.

Techniques for Imputation:

- **SimpleImputer:**
  - *Mean*: Replaces missing values with the mean of the feature.
  - *Median*: Replaces missing values with the median of the feature.
  - *Most Frequent*: Replaces missing values with the most frequent value in the feature.

- **Linear Regression:**
  This approach uses other features to predict the missing values based on a linear relationship.

- **K-Nearest Neighbors (KNN):**
  KNN uses a specified number of closest neighbors to estimate missing values based on their similarities.

- **MICE (Multivariate Imputation by Chained Equations):**
  MICE involves modeling each feature with missing values as a function of other features, iterating until convergence.

Project Process:

We began by creating a script to remove random values from our dataset, ensuring that missing values were spread across various columns. After this, we applied each imputation technique and evaluated the results.

- **Evaluation Metrics:**
  - Visual inspection of the data distribution before and after imputation**

- **Results:**
  The results of our analysis are summarized below:

  - SimpleImputer (Mean): This technique worked well for numerical columns but might not always maintain the original data distribution.

  - SimpleImputer (Median): Median imputation was robust to outliers and generally resulted in better predictions for skewed distributions.

  - SimpleImputer (Most Frequent): This strategy worked best for categorical data and features with a high number of repeated values.

  - Linear Regression: Provided accurate results but could be sensitive to outliers or complex relationships in the data and more.
  - KNN: This method was not effective for our dataset because most of the time it's imputed the values exactly similar to mean.

- MICE: The most sophisticated approach, providing reliable imputations by considering relationships between multiple features, It worked well for numerical data but it is more complex and expensive to use.

DF ROW NO.	Column	ORIGINAL VALUE	PREDICTED VALUE	IMPUTATION TECHNIQUE	REASONS
150	wheelbase	96.9	97	Median	We used median because data distribution is skewed and has outliers.
2	carlength	168.8	167.3	MICE through LinearReggression	We used MICE because it's closer to actual value but we have to consider that it took 4 seconds to compute the value over 50 iterations. but we can also go for a mean value that is 174 because data is symmetrically distributed (normal distribution).
37	carlength	157.1	157.3	Most Frequent	it's closer to actual value but as mentioned above, we can impute the mean and MICE also.
85	carwidth	66.3	65.905882	Mean	it's more closer to actual value and data distribution is also normal
11	carheight	54.3	65.5	Median	We used median because the iterative method also gave approx same value which is still more than the actual, and data distribution is also not normal and has outliers.
68	curbweight	3515	3518	LinearRegression	Used this because its value is closer to actual, because all statistical techniques are way less than the actual but we have to consider that it's more complex and expensive when used in production.
102	curbweight	3095	3045	LinearRegression	Used this because its value is closer to actual, because all statistical techniques are way more lesser than the actual but we have to consider that it's more complex and expensive when used in production.
154	curbweight	2280	2275	LinearRegression	Used this because its value is closer to actual, because all statistical techniques are way less than the actual but we have to consider that it's more complex and expensive when used in production.
109	enginesize	152	147.0264094	LinearRegression	Used this because its value is closer to actual, because all statistical techniques are way less than the actual but we have to consider that it's more complex and expensive when used in production.
159	boreratio	3.27	3.31	Median	Although all techniques give approx values, that's why using statistical techniques is more beneficial, but choose median specifically because data is skewed.
101	compression ratio	8.5	9	Median	Although all techniques give approx values, that's why using statistical techniques is more beneficial, but choose median specifically because

					data is skewed.
125	compression ratio	7	9	Median	Although all techniques give approx values, that's why using statistical techniques is more beneficial, but choose median specifically because data is skewed.
25	horsepower	68	68	Most Frequent	it's closer to actual value but as mentioned above, we can impute with MICE also which is 69 but it takes time to compute and is expensive for production purposes.
93	horsepower	69	68	Most Frequent	it's closer to actual value but as mentioned above, we can impute with MICE also which is 69 but it takes time to compute and is expensive for production purposes.
115	peakrpm	4151	4182.167962	MICE through LinearReggression	it's more closer to actual value but we can also to with mean that is 5129 because it cheaper and data distribution is also normal
91	citympg	45	43.104	LinearRegression	it's giving closer to actual value rest are for away to actual value
188	highwaympg	42	43.90745587	LinearRegression	it's giving closer to actual value rest are for away to actual value
5	price	17450	17433.99401	LinearRegression	Although imputed with Linear Regression still less accurate, else all other techniques went in haywire.
109	price	13200	16660.62116	LinearRegression	Although imputed with Linear Regression still less accurate, else all other techniques went in haywire.
199	price	18420	15223.73758	LinearRegression	Although imputed with Linear Regression still less accurate, else all other techniques went in haywire.

Self-Learning and Industry Uses:

- **Self-Learning:**
  - Learned that imputation techniques vary in complexity and applicability depending on the dataset.
  - Realized the importance of understanding the data structure to choose the best imputation method.
  - Gained experience in implementing various imputation techniques and analyzing their impact on data accuracy.

- **Industry Uses:**
  - Healthcare: Used for filling missing patient information in medical records.
  - Finance: Applied to maintain data integrity in stock market and financial analysis.

- Marketing: Helps in accurate customer segmentation and profiling despite incomplete data.
- Automotive: Critical for maintaining complete datasets for car prices, insurance, and maintenance.

- **Applications:**
  - Data Cleaning: The imputation techniques can be applied to improve data quality and consistency in any dataset with missing values.
  - Predictive Modeling: Handling missing data allows for more accurate machine learning models in various domains.
  - Data Analysis: Imputation methods are used to ensure complete datasets for effective data analysis and reporting.

- **Conclusion:**
  - Imputation techniques play a significant role in dealing with missing values in datasets, with different methods offering varying levels of complexity and accuracy.
  - SimpleImputer is a good baseline, while Linear Regression, KNN, and MICE provide advanced options for more complex scenarios.
  - The choice of imputation technique should align with the dataset's characteristics and the specific requirements of the analysis or project.
  - Our project demonstrated that combining multiple imputation techniques can lead to improved accuracy and data integrity.

- **Contributions:**
  - **Pankaj**:
    - Contributed to the implementation and analysis of the SimpleImputer methods.
    - Conducted research on industry uses of imputation techniques.
    - Helped with the evaluation and comparison of different imputation approaches.

  - **Deepak**:
    - Focused on the implementation and analysis of Linear Regression, KNN, and MICE techniques.
    - Conducted research on self-learning aspects of imputation techniques.
    - Played a key role in compiling the results and drawing conclusions from the analysis.

Together, we discussed the results, compared the techniques, and wrote the report. The collaboration allowed us to learn from each other and gain a deeper understanding of how different imputation methods work in practice.