Feature Engineering - Handling Missing Values

By Binoy Patra

**Introduction:**

**Missing values** are instances where data points that should be present in a dataset are either not recorded or unavailable. They occur when no data value is stored for a variable in an observation.

**Types of Missing Data**

1. **Missing Completely at Random (MCAR)**:
   * The missingness is entirely random and unrelated to any observed or unobserved data.
   * Example: A survey respondent skips a question due to a random error.
2. **Missing at Random (MAR)**:
   * The missingness depends on observed data but not on the missing data itself.
   * Example: Higher-income individuals are less likely to report their income, but their education level (observed data) can predict the missingness.
3. **Missing Not at Random (MNAR)**:
   * The missingness depends on the unobserved (missing) data itself.
   * Example: People with very high incomes are less likely to disclose their income.

**Common Causes of Missing Data**

* Data entry errors.
* Non-response in surveys or questionnaires.
* Equipment failures during data collection.
* Data corruption or loss.
* Intentional omissions (e.g., sensitive questions skipped by respondents).

**Impact of Missing Values**

* Reduces the completeness and reliability of the dataset.
* Affects statistical analysis and model performance.
* Can lead to biased conclusions if not handled appropriately.

**Why Handle Missing Values?**

1. **Avoid Bias**: Missing values can lead to biased results, especially if the missingness is not random (e.g., MAR or MNAR).
2. **Improve Model Performance**: Most machine learning algorithms cannot handle missing data, and proper handling ensures better predictions and accuracy.
3. **Preserve Data Integrity**: Missing values can distort statistical measures like mean, variance, and correlations, leading to incorrect conclusions.
4. **Maximize Data Utilization**: Handling missing values prevents significant data loss from deletion methods, ensuring the dataset retains its full potential.
5. **Ensure Analytical Compatibility**: Many statistical and machine learning methods require complete data, so addressing missing values is essential for their proper functioning.

**How To Handle Missing Values?**

**1. Deletion Methods**

**Description**: Remove rows or columns with missing values.

* **Pros**: Simple, no assumptions required.
* **Cons**: Loss of data, reduced sample size.
* **When to apply**: When missing data is minimal (<5%) and randomly distributed (MCAR).

**2. Imputation Methods**

**Mean/Median/Mode Imputation**

**Description**: Replace missing values with the mean (for both normal and skewed distribution) , median (for skewed distribution), or mode (for categorical variable).

* **Pros**: Easy to implement, fast.
* **Cons**: Distorts data distribution, ignores variability, introduce outliers.
* **When to apply**: When missing data is small and data distribution is uniform.

**Forward/Backward Filling**

**Description**: Fill missing values with the previous (forward) or next (backward) available value.

* **Pros**: Works well for time-series data.
* **Cons**: Assumes continuity, not suitable for abrupt changes.
* **When to apply**: For sequential or time-series datasets.

**Arbitrary Value Imputation**

**Description**: Replace missing values with a fixed, arbitrary number (e.g., -999/1/0 in case of numerical variable) or values (NA in case of categorical variable).

* **Pros**: Simple, preserves missingness as identifiable.
* **Cons**: Can distort relationships and metrics, finding perfect number/text.
* **When to apply**: When missing values are meaningful or require clear identification.

**Linear Interpolation**

**Description**: Estimate missing values by interpolating between known values.

* **Pros**: Maintains data trends.
* **Cons**: Assumes linear relationships, unsuitable for non-linear data.
* **When to apply**: For continuous data with predictable trends.

**End of Distribution Imputation (Applicable for Numeric variable)**

**Description**: Replace missing values with extreme values (e.g., min or max).

* **Pros**: Captures uniqueness of missingness.
* **Cons**: Can distort metrics and relationships.
* **When to apply**: For outlier detection or when missing values indicate extreme cases.

**Multiple Imputation**

**Description**: Generate multiple plausible imputed datasets and combine results.

* **Pros**: Accounts for uncertainty, statistically robust.
* **Cons**: Computationally intensive, requires expertise.
* **When to apply**: For large datasets with MAR or MCAR missingness.

**Regression Imputation**

**Description**: Predict missing values using a regression model.

* **Pros**: Utilizes relationships among variables.
* **Cons**: Assumes linearity, may introduce bias.
* **When to apply**: When strong relationships exist between variables.

**3. Model-Based Methods**

**Decision Trees (e.g., Random Forest)**

**Description**: Use decision trees to predict and impute missing values.

* **Pros**: Captures non-linear relationships, works for mixed data types.
* **Cons**: Computationally expensive, risk of overfitting.
* **When to apply**: For complex datasets with non-linear dependencies.

**k-Nearest Neighbours (KNN)**

**Description**: Impute values using the mean/mode of similar data points (neighbours).

* **Pros**: Flexible, works well with structured data.
* **Cons**: Computationally heavy for large datasets.
* **When to apply**: When clear local patterns exist in the data.

**4. Other Methods**

**Using Indicator Variables for Missingness**

**Description**: Add binary columns to indicate missing values.

* **Pros**: Retains missingness information, prevents data loss.
* **Cons**: Increases dataset complexity, may overfit.
* **When to apply**: When missingness itself is meaningful or informative (e.g., MNAR scenarios).

**Things to keep in mind before and after handling missing values:**

Handling missing values requires careful consideration before and after applying techniques to ensure data quality and integrity.

**Before Handling Missing Values**

**1. Understand the Missingness Mechanism (MCAR/MAR/MNAR)**

**2. Assess the Extent of Missingness**

* **Percentage of missing values**: Columns with >50% missingness may be better dropped unless critical.
* **Visualize missingness**: Use heatmaps or bar plots to understand patterns.

**3. Consider the Data Context**

* Consult domain knowledge to understand the importance of features and possible reasons for missingness.
* Identify if missing values hold meaning (e.g., "NA" indicating 'Not Applicable').

**4. Analyze the Impact of Missing Data**

* Evaluate correlations and relationships between missing values and target variables.
* Identify if imputing missing data can lead to biased results.

**5. Separate Training and Testing Data**

* Never impute on the entire dataset; always split into train-test first to avoid data leakage.

**During Handling Missing Values**

**6. Choose the Right Method**

* For **low missingness** (<5%): Simple imputation (mean/median/mode) or deletion may suffice.
* For **high missingness** (>50%): Consider advanced methods like multiple imputations or domain-specific techniques.
* For **categorical variables**: Mode imputation or adding indicator variables works well.
* For **time-series data**: Use forward/backward filling or interpolation.

**7. Ensure Consistency**

* Use the same imputation logic for both training and test datasets.
* Document your chosen method and its rationale for reproducibility.

**After Handling Missing Values**

**8. Verify the Impact**

* **Check data quality**: Ensure no missing values remain after imputation.
* **Distribution analysis**: Compare the distribution of imputed data with the original data.

**9. Validate Relationships**

* Check if correlations or patterns are preserved post-imputation.
* Use scatter plots or heatmaps to verify relationships.

**10. Evaluate Model Performance**

* Compare model results with and without imputation to ensure the handling method improves performance.
* For complex datasets, use cross-validation to ensure robust evaluation.

**11. Document Your Process**

* Record the methods used, assumptions made, and why specific techniques were chosen.
* Helps in troubleshooting and improving reproducibility.

**12. Be Aware of Overfitting**

* Advanced imputation techniques (e.g., KNN or regression) may overfit the training data.
* Use separate imputation steps for training and testing data.

**Common Pitfalls to Avoid**

* **Blindly dropping data**: Avoid dropping rows or columns without evaluating their importance.
* **Over-imputation**: Imputing too much can distort data variability.
* **Ignoring missingness patterns**: Overlooking patterns can lead to incorrect assumptions (e.g., MAR vs MNAR).
* **Failing to validate**: Always verify imputation results for logical consistency.

By keeping these points in mind, you'll handle missing values effectively while maintaining data integrity and improving the quality of your analysis or model.