Feature Engineering - Handling Duplicate Values

By Binoy Patra

**Introduction:**

Duplicate values are rows in a dataset that have identical data in all or specific columns. These entries repeat the same information and do not contribute any new insights.

**Why Are They in the Dataset?**

1. **Data Collection Errors**: Issues during data gathering, such as repeated submissions or system glitches.
2. **Data Merging**: Combining datasets without proper cleaning can result in duplicate rows.
3. **User Actions**: Multiple entries for the same user or event (e.g., resubmissions in a form).
4. **Crawling/Scraping Errors**: Data scraping might capture the same information multiple times.

**Effects of Duplicate Values**

1. **Distorted Analysis**: Can lead to overestimation of trends or metrics.
2. **Skewed Model Training**: Machine learning models can overfit or misinterpret patterns due to redundant data.
3. **Resource Waste**: Larger datasets increase computational cost and storage requirements unnecessarily.
4. **Misleading Insights**: Reports and dashboards may present inaccurate results.

**Why Remove Duplicate Values?**

1. **Improve Data Quality**: Ensure the dataset is clean and represents unique observations.
2. **Accurate Insights**: Maintain the integrity of analysis and reporting.
3. **Optimize Performance**: Reduce the dataset size, speeding up processing and analysis.
4. **Avoid Bias in Models**: Prevent duplicate entries from disproportionately influencing model predictions.

Here is the step-by-step **process** for handling duplicate values:

1. **Identify Duplicates**:
   * Analyze the dataset to check if duplicate values exist. This includes examining rows or specific columns to determine duplicate patterns.
2. **Understand the Context**:
   * Investigate why duplicates exist. Are they due to errors, merging issues, or legitimate repeated data?
   * Assess whether they are full-row duplicates or duplicates based on certain columns.
3. **Decide on Retention**:
   * Determine if duplicates should be removed or retained. For example:
     + Retain the first occurrence.
     + Retain the last occurrence.
     + Retain duplicates if they add value (e.g., time-series or grouped data).
4. **Remove or Resolve Duplicates**:
   * If needed, remove duplicate rows or partial duplicates based on the context and analysis. Use appropriate methods to clean the data.
5. **Verify Cleaning**:
   * Re-check the dataset to confirm all unwanted duplicates are removed and no valuable data is lost.
6. **Document Changes**:
   * Record the number of duplicates found and removed, along with the criteria used for cleaning, to maintain transparency in the data-cleaning process.