* New Category

**Goal : Discover new categories in the test set.**

* Datasets Size Comparison

**Goal: Verify the size of the test dataset by comparing it to the training dataset.**

* Date Train Test Validation Leakage Duplicates

Any record that mistakenly shares data with another record in a Database is considered duplicate data. It is simple to identify duplicate data, which typically occurs during the transmission of data between systems. The most common instance of duplicate data is a carbon copy of a record.

**Goal: Inspect train data for the presence of test dates.**

* Date Train Test Validation Leakage Overlap

**Goal: Examine test data that is older than the most recent date in train.**

* Dominant Frequency Change

Dominant Frequency Change is a data integrity test that examines if dominant values have considerably increased between testing data and training data. Sharp changes in dominating values may signal a fault with the data collection or data processing pipeline (for instance, a sharp spike in a common null or constant value), and will prevent the model from generalizing effectively. This check aims to identify these flaws early in the production process.

This test compares the dominating values of each feature in the testing dataset to the corresponding values in the training dataset. If the ratio of testing to training dominating values exceeds a certain level, the test fails.

The Characteristics of a Dominant Value

The dominant value is defined as a value that is at least the dominance ratio times more frequent than the next most frequent value. The check's dominance ratio is a customizable parameter.

Goal: Determine whether dominant values have increased significantly between the test data and the reference data.

* Feature Label Correlation Change
* Index Leakage
* New Label
* String Mismatch Comparison
* Train Test Feature Drift
* Train Test Label Drift
* Train Test Samples Mix
* Whole Dataset Drift