* 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

Goal: Return the Predictive Power Score for each feature to measure its ability to predict the label.

* Index Leakage

Goal : Determine if test indexes exist in train data.

* New Label

Goal: Discover new labels for the test.

* String Mismatch Comparison

Why is the check being issued?

The check examines the same category column in training and testing to determine if there are variants of similar strings that exist in testing but not in training. Identifying these discrepancies is useful for preventing errors when inferring from testing data. For instance, the category 'New York' exists in both training dataset and our testing dataset. We would like to be notified when the testing results include a new variety of the training dataset, so that we may resolve the issue.

What Does String Mismatch Mean?

To identify string mismatch, we reduce each string to its fundamental form. The base form is the string with only lowercase alphabetic characters. (The root form of "Cat-9?!" is "cat9") If two strings have the identical base form, they are termed identical.

Goal: Identify distinct string category variants within the same categorical column of two datasets.

* Train Test Feature Drift

What is characteristic drift?

Drift is essentially a change in the distribution of data over time, and it is one of the primary reasons why the performance of machine learning models diminishes with time.

Feature drift is a type of data drift that affects a specific feature of a dataset.

This testing detects feature drift by applying univariate measurements to each feature column individually. A domain classifier, which is utilized in the Whole Dataset Drift check, is an additional way for detecting drift.

Using statistical techniques, calculate the drift between the training data and the testing dataset for each feature.

* Train Test Label Drift

Drift is essentially a change in the distribution of data over time, and it is one of the primary reasons why the performance of machine learning models diminishes with time.

Label drift happens when the label itself drifts.

Goal: Using statistical metrics, compute label drift between the training and testing datasets.

* Train Test Samples Mix

Samples mix occurs when the training and testing datasets contain overlapping samples. We utilize the testing dataset to evaluate the performance of our model, and having samples in common with the training data will result in skewed metrics that do not reflect the actual performance we will achieve in the real world. Therefore, we must always avoid sample mixing.

Goal: Observing data in the testing data that also exist in the training data.

* Whole Dataset Drift

What Is the Meaning of Multivariate Drift?

Drift is essentially a change in the distribution of data over time, and it is one of the primary reasons why the performance of machine learning models diminishes with time.

A multivariate drift is a drift that affects the relationships between many features and cannot be detected by univariate drift detection methods. The entire dataset drift check attempts to identify multivariate drift between two input datasets.

Using a domain classifier, this check finds multivariable drift. Other techniques for detecting drift include univariate measurements, which are employed by other tests, such as the Train Test Feature Drift check.

**Goal : Using a model trained to differentiate between the training and testing sets datasets, compute the drift between the whole training and testing sets datasets.**