Mercury

Mercury.robust

robust test

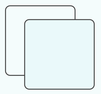
BBVA

mercury-robust is a framework for performing robust testing of models and/or datasets.

It provides a series of predefined and configurable tests cases to ensure the robustness of your ML pipelines. For instance, you can quickly verify whether a model discriminates against a group or check if re-training the model with the same data yields the exact predictions.

DataTest

Same Schema



Ensures that the DataFrame has the same columns and feature types as those specified in the DataSchema.

Data Drift

Checks that the individual feature distributions have not significantly changed between a reference DataSchema and a pandas.DataFrame.

Linear Combinations

Checks that you have no redundant or unnecessary columns in your pandas.Dataframe.

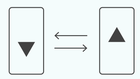
No Duplicates

Checks that you don't have repeated samples in your dataset, which can add bias on your performance metrics.

Noisy Labels

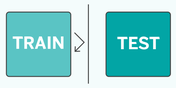
Examines the labels of your dataset to guarantee they have a minimum quality. We consider labels low-quality when there are many incorrectly labeled samples or when the separation between labels is not evident.

Cohort Performance



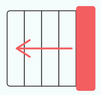
This test compares a particular metric (e.g. 'accuracy') between several groups specified by a categorical variable in your pandas.DataFrame ('group_col').

Sample Leaking



Checks if the test or validation dataset contains samples that are already included in the training dataset.

Label Leaking



Checks that you don't have any feature leaking information about the target variable.

ModelTest

Model Simplicity

Compares the performance of your model to a simpler baseline (by default a linear model, although you can specify your custom baseline).

Model Reproducibility

Trains a model twice and checks that the two versions' predictions (or a particular metric) are not too different.

Drift Metric Resistance

Adds artificial drift to a reference dataset and tests your model on it. If a chosen metric (e.g. 'accuracy') changes, the test will fail, indicating that your model is weak against drift.

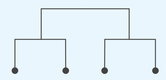
Drift Predictions Resistance

Adds artificial drift to a reference dataset and tests your model on it. If lots of predictions change the test will fail, indicating that your model is weak against drift.

Feature Checker

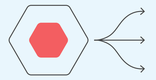
Estimates the importance of your features and retrains your model by removing the least important ones one by one.

Tree Coverage



Works only with tree-based models (mainly those of scikit-learn). It checks that, given a test dataset and a trained model, the samples "activate" a minimum amount of branches in their tree(s).

Classification Invariance



Verifies that the model's prediction remains unchanged when a perturbation is applied that should not impact the label of the samples.

TEST SUITE

Holds a set of tests so you only have to run them once