gitflow_training_pipeline_data_drift_detector

The suite is composed of various checks such as: Feature Label Correlation Change, Date Train Test Leakage Overlap, Train Test Samples Mix, etc...
Each check may contain conditions (which will result in pass / fail / warning / error , represented by ✓ / ✖ / ! / ?!) as well as other outputs such as plots or tables.
Suites, checks and conditions can all be modified. Read more about custom suites.

Conditions Summary

Status	Check	Condition	More Info
✓	Category Mismatch Train Test	Ratio of samples with a new category is less or equal to 0%	Passed for 2 relevant columns
✓	Datasets Size Comparison	Test-Train size ratio is greater than 0.01	Test-Train size ratio is 0.25
✓	Feature Label Correlation Change	Train features' Predictive Power Score is less than 1	Passed for 31 relevant columns
✓	New Label Train Test	Number of new label values is less or equal to 0	No new labels found
✓	String Mismatch Comparison	No new variants allowed in test data	No relevant columns to check were found
✓	Train Test Feature Drift	categorical drift score < 0.2 and numerical drift score < 0.1	Passed for 31 columns out of 31 columns. Found column "worst texture" has the highest numerical drift score: 0.06
✓	Train Test Label Drift	categorical drift score < 0.2 and numerical drift score < 0.1 for label drift	Label's drift score Cramer's V is 0
✓	Train Test Samples Mix	Percentage of test data samples that appear in train data is less or equal to 10%	No samples mix found
✓	Whole Dataset Drift	Drift value is less than 0.25	Found drift value of: 0.05, corresponding to a domain classifier AUC of: 0.52

Check With Conditions Output

Datasets Size Comparison

Verify test dataset size comparing it to the train dataset size. Read More...

Conditions Summary

Status	Condition	More Info
✓	Test-Train size ratio is greater than 0.01	Test-Train size ratio is 0.25

Additional Outputs

	Train	Test
0	455	114

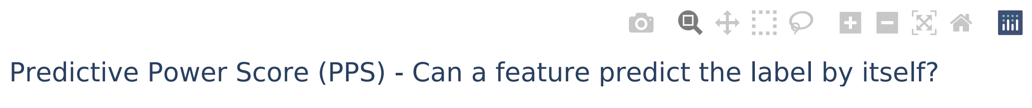
Feature Label Correlation Change

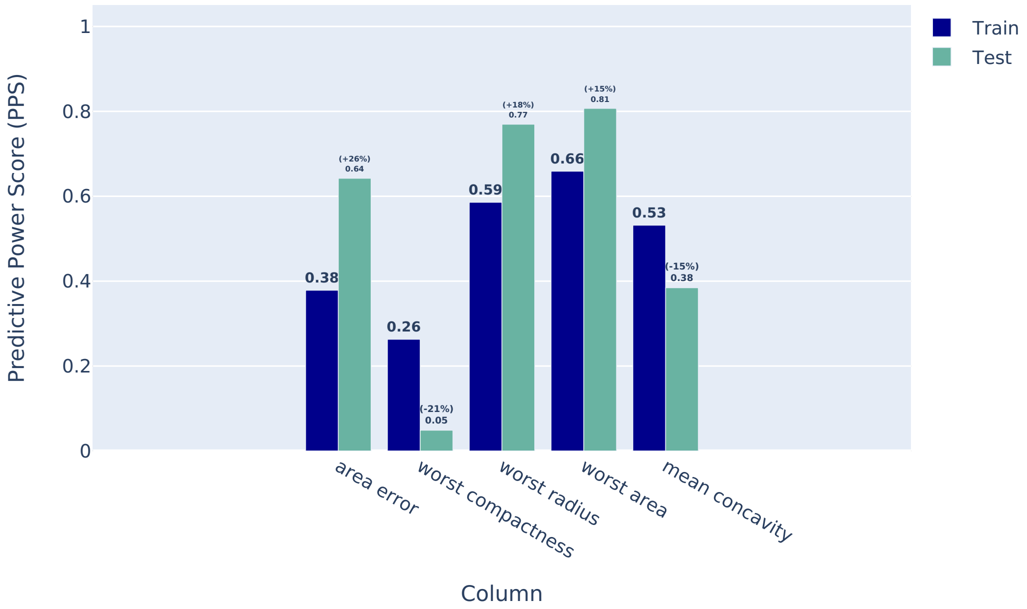
Return the Predictive Power Score of all features, in order to estimate each feature's ability to predict the label. Read More...

Conditions Summary

Status	Condition	More Info
✓	Train features' Predictive Power Score is less than 1	Passed for 31 relevant columns

Additional Outputs





The Predictive Power Score (PPS) is used to estimate the ability of a feature to predict the label by itself. (Read more about Predictive Power Score)

In the graph above, we should suspect we have problems in our data if:

1. **Train dataset PPS values are high**:
Can indicate that this feature's success in predicting the label is actually due to data leakage,
meaning that the feature holds information that is based on the label to begin with.
2. **Large difference between train and test PPS** (train PPS is larger):
An even more powerful indication of data leakage, as a feature that was powerful in train but not in test
can be explained by leakage in train that is not relevant to a new dataset.
3. **Large difference between test and train PPS** (test PPS is larger):
An anomalous value, could indicate drift in test dataset that caused a coincidental correlation to the target label.

Train Test Feature Drift

Calculate drift between train dataset and test dataset per feature, using statistical measures. Read More...

Conditions Summary

Status	Condition	More Info
✓	categorical drift score < 0.2 and numerical drift score < 0.1	Passed for 31 columns out of 31 columns. Found column "worst texture" has the highest numerical drift score: 0.06

Additional Outputs

The Drift score is a measure for the difference between two distributions, in this check - the test and train distributions.
The check shows the drift score and distributions for the features, sorted by drift score and showing only the top 5 features, according to drift score.
If available, the plot titles also show the feature importance (FI) rank.

Train Test Label Drift

Calculate label drift between train dataset and test dataset, using statistical measures. Read More...

Conditions Summary

Status	Condition	More Info
✓	categorical drift score < 0.2 and numerical drift score < 0.1 for label drift	Label's drift score Cramer's V is 0

Additional Outputs

The Drift score is a measure for the difference between two distributions, in this check - the test and train distributions.
The check shows the drift score and distributions for the label.

Check Without Conditions Output

Other Checks That Weren't Displayed

Check	Reason
Date Train Test Leakage Duplicates	Dataset does not contain a datetime
Date Train Test Leakage Overlap	Dataset does not contain a datetime
Index Train Test Leakage	Dataset does not contain an index
Category Mismatch Train Test	Nothing found
Dominant Frequency Change	Nothing found
New Label Train Test	Nothing found
String Mismatch Comparison	Nothing found
Train Test Samples Mix	Nothing found
Whole Dataset Drift	Nothing found