gitflow_training_pipeline_data_integrity_checker

The suite is composed of various checks such as: Feature Label Correlation, Special Characters, Columns Info, 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
✓	Conflicting Labels	Ambiguous sample ratio is less or equal to 0%	Ratio of samples with conflicting labels: 0%
✓	Data Duplicates	Duplicate data ratio is less or equal to 0%	Found 0% duplicate data
✓	Feature Label Correlation	Features' Predictive Power Score is less than 0.8	Passed for 31 relevant columns
✓	Single Value in Column	Does not contain only a single value	Passed for 32 relevant columns
✓	Mixed Data Types	Rare data types in column are either more than 10% or less than 1% of the data	32 columns passed: found 0 columns with negligible types mix, and 32 columns without any types mix
✓	Mixed Nulls	Number of different null types is less or equal to 1	Passed for 32 relevant columns
✓	Special Characters	Ratio of samples containing solely special character is less or equal to 0.1%	Passed for 32 relevant columns
✓	String Length Out Of Bounds	Ratio of string length outliers is less or equal to 0%	No relevant columns to check were found
✓	String Mismatch	No string variants	No relevant columns to check were found

---

Check With Conditions Output

Feature Label Correlation

Return the PPS (Predictive Power Score) of all features in relation to the label. Read More...

Conditions Summary

Status	Condition	More Info
✓	Features' Predictive Power Score is less than 0.8	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). A high PPS (close to 1) can mean 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.

Check Without Conditions Output

Columns Info

Return the role and logical type of each column. Read More...

Additional Outputs

* showing only the top 10 columns, you can change it using n_top_columns param

	target	Unnamed: 0	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points
0	label	numerical feature	numerical feature	numerical feature	numerical feature	numerical feature	numerical feature	numerical feature	numerical feature	numerical feature

Outlier Sample Detection

Detects outliers in a dataset using the LoOP algorithm. Read More...

Additional Outputs

The Outlier Probability Score is calculated by the LoOP algorithm which measures the local deviation of density of a given sample with respect to its neighbors. These outlier scores are directly interpretable as a probability of an object being an outlier (see link for more information).

	Outlier Probability Score	Unnamed: 0	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error	perimeter error	area error	smoothness error	compactn error
0	0.84	213	17.42	25.56	114.50	948.00	0.10	0.11	0.17	0.07	0.13	0.06	0.53	1.67	3.77	58.53	0.03	0.09
1	0.80	288	11.26	19.96	73.72	394.10	0.08	0.12	0.09	0.06	0.26	0.06	0.49	1.91	2.88	34.68	0.02	0.08
2	0.78	212	28.11	18.47	188.50	2499.00	0.11	0.15	0.32	0.16	0.16	0.06	2.87	1.48	21.98	525.60	0.01	0.03
3	0.78	491	17.85	13.23	114.60	992.10	0.08	0.06	0.04	0.04	0.12	0.05	0.48	1.05	3.16	50.95	0.00	0.01
4	0.73	192	9.72	18.22	60.73	288.10	0.07	0.02	0.00	0.00	0.17	0.06	0.35	4.88	2.23	21.69	0.00	0.01

Other Checks That Weren't Displayed

Check	Reason
Feature Feature Correlation - Train Dataset	module 'numpy' has no attribute 'bool'
Identifier Leakage - Train Dataset	Dataset does not contain an index or a datetime
Conflicting Labels	Nothing found
Data Duplicates	Nothing found
Single Value in Column	Nothing found

Mixed Data Types	Nothing found
Mixed Nulls	Nothing found
Special Characters	Nothing found
String Length Out Of Bounds	Nothing found
String Mismatch	Nothing found