BigML Anomaly Cheat Sheet

Anomaly Configuration

Anomaly Configuration Options

Option	Description	Default	API Name
Number of anomalies	Sets the number of instances with top anomaly scores to be displayed in the anomalies view. You can request up to 1,024 anomalies.	10	top_n
Forest size	Sets the number of decision trees used by the anomaly detector. This must be a number up to 256 (1,000 if you are using the BigML API). Higher numbers tend to give better results although they take longer to process.	128	forest_size
Constraints	Makes the trees more sensitive to anomalies. This option tends to inflate the anomaly scores (and it is more costly in terms of computational costs), but it can also make the trees more effective at flagging anomalous data, especially with categorical data.	False	constraints
ID fields	Fields not used to compute anomalies, but their values are included in the DATA INSPECTOR in the anomalies view. Non-preferred fields are not eligible as ID fields. If you want to include a non-preferred field as an ID field, you will need to first set that field as preferred.	[ ]	id_fields

Sampling

Option	Description	Default	API Name
Rate	Sets the proportion of the dataset you want to consider between 0% and 100%.	100%	sample_rate
Range	Specifies a subset of instances from which to sample, e.g., from instance 5 to instance 1,000. The **Rate** you set will be computed over the **Range** configured.	(1, max. rows in dataset)	range
Sampling	Allows you to choose between a random sampling or a deterministic sampling. When using deterministic sampling the random-number generator will always use the same seed, producing repeatable results.	Random	seed
Replacement	Allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance cannot be selected more than once.	False	replacement
Out of bag	Selects only the out-of-bag instances for the currently defined sample. If an instance is not selected as part of a sample, it is considered out of bag. It is only selectable when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag

Batch Score Configuration

Output File Options

Option	Description	Default	API Name
Fields separator	Allows you to choose the best separator for your fields.	Comma	separator
Show/hide fields	Allows you to show or hide the rest of the fields in your output file.	True	output_fields
Headers	Allows you to show or hide the names of your columns in the output file.	True	header
Score column name	Allows you to set the name for the anomaly score column in your output file.	Score	score_name
Fields importance	Allows you to include the relative importances of each input field in the anomaly score per instance.	False	importance
New line	Sets the character to use as the line break in the generated csv file: “LF”, “CRLF”.	LF	newline

Output Dataset

Option	Description	Default	API Name
Output dataset	Defines whether a dataset with the results should be automatically created or not.	True	output_dataset