BigML Model Cheatsheet

Model configuration

Option	Description	Default	API Name
Objective field	Selects the field you want to predict. It can be a categorical or numeric field.	Last valid field in dataset	objective_field
Pruning	Allows you to choose one of the three pruning strategies BigML offers. Smart Pruning only considers pruning nodes with less than 1% support. Statistical Pruning applies pruning to all the nodes. The last option is No pruning.	Smart Pruning	stat_pruning & selective_pruning
Missing splits	Tells whether to consider missing data as a split criterion.	False	missing_splits
Node threshold	Defines the maximum number of computed nodes for a model. When the number of computed nodes is greater than this threshold, model growth stops. You can set a value between 3 and 2,000.	512	node_threshold

Weight field options

Option	Description	Default	API Name
Balance objective	Sets instance weights so that each class has equal influence on the model. This is only available for classification models.	False	balance_objective
Weight field	Sets instance weights using the values of the given field. The selected field must be numerical and it must not contain missing values. This is valid for both regression and classification models.	False	weight_field
Objective weights	Sets a specific weight for each class of the objective field. If a class is not listed, it is assumed to have a weight of 1. Weights of 0 are also valid. This option is only available for classification models.	False	objective_weights

Sampling options

Option	Description	Default	API Name
Rate	Sets the proportion of the dataset to be sampled 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 sampling, it is considered an out-of-bag instance. It is only selectable when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag

Ordering options

Option	Description	Default	API Name
Deterministic shuffling	Ensures the row shuffling of a dataset is always the same, so that evaluating a model from the same dataset always yields the same results.	True	ordering:0
Linear	Selects the instances in the order they are listed to build the model. If you know that your instances are already in random order, set the shuffling to linear so that the model will be constructed faster.	False	ordering:1
Random shuffling	Takes a different sampling each time you build your model.	False	ordering:2

Prediction configuration

Missing strategy options

Option	Description	Default	API Name
Last prediction	Specifies that when a missing value is found in the testing data for a decision node, the prediction will be that from the parent of the missing split.	True	missing_strategy:0
Proportional missing strategy	Specifies that when a missing value is found in the input data for a decision node, the prediction is based on all the subtrees of a missing split. This recombines their predictions based on the proportion of data in each subtree.	False	missing_strategy:1

Confidence or Probability threshold

Option	Description	Default	API Name
Confidence or probability threshold	A confidence or probability percentage between 0% and 100% that can be used with classification ensembles so that they only return the positive class when the confidence or probability on the prediction is above the established threshold.	Null	operating_point

Default Numeric Values

Option	Description	Default	API Name
Default numeric value	Replaces missing numeric values in your dataset by the field's maximum, mean, median, minimum, or zero.	Null	default_numeric_value

Output file options

Option	Description	Default	API Name
Fields separator	Allows you to choose the best separator for your fields.	Comma	separator
New line	Sets the character to use as the line break in the generated csv file: “LF”, “CRLF”.	LF	newline
Show/hide fields	Shows or hides the rest of the fields in your output file.	True	output_fields
Headers	Shows or hides the names of your columns in the output file.	True	header
Prediction column name	Sets the name you want for the objective field. By default BigML takes the name of the model’s objective field.	Objective field name	prediction_name
Include confidence	Includes an additional column with the confidence (or expected error) per instance associated with the predictions.	False	confidence
Confidence column name	Sets the name you want for the confidence (or expected error) field.	Confidence	confidence _name
Include probability	Includes an additional column with the probability of the predicted class.	False	probability
Probability column name	Sets the name you want for the probability column. By default it is named “probability”.	Confidence	probability_name
Confidences	Includes a column for each of the objective field classes indicating their confidences per instance predicted. This will add a column per field, named “<objective_field_class> confidence”.	False	confidences
Probabilities	Includes a column for each of the objective field classes indicating their probabilities per instance predicted. This will add a column per field, named “<objective_field_class> probability”.	False	probabilities
Importances	Defines whether the batch prediction includes a column for each of the field importances for the model predictions. There is a column per field, named "<field_name> importance".	False	importance

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