BigML Ensemble Cheatsheet

Ensemble configuration

Ensemble configuration options

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
Type	Allows you to choose which algorithm to use to build the ensemble. Decision Forests use a random sample of instances when building each model and/or a random subset of the input data fields when generating splitting rules. Boosted trees or gradient boosting trees iterativly build each model trying to correct the mistakes made by the previous model	Decision Forests	boosting
Number of models	Sets the total number of models in the ensemble when building Decision Forests ensembles. Valid values lie between 2 and 1,000 models. For Boosted Trees the maximum number of models is 2,000.	10	number_of _models
Number of iterations	Sets the total number of iterations in the ensemble when building Boosted trees. For regression ensembles one model is built for each iterations, for classification enembles N models are built in each iteration letting N be the total number of classes in the objective field. Valid values lie between 2 and 1,000 iterations with a limit of 2,000 models.	10	number_of _iterations

Trees options

Option	Description	Default	API Name
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
Randomize	Allows a randomly chosen fields to be considered at each split. If Decision Forests is selected it createase a Random Decision Forests. By default BigML takes the squared root of the total number of fields. You can set a fixed value or a ratio.	False	randomize
Random candidates	Sets the number of randomly chosen fields to be considered at each split for Random Decision Forests. By default BigML takes the squared root of the total number of fields. You can set a fixed value or a ratio.	$\sqrt{n}$	random _candidates & random _candidate_ratio

Boosting options

Option	Description	Default	API Name
Early stopping	Tries to find the optimal number of iterations by testing the single models after every iteration and resulting in an early stop if not significant improvement is made. You can select the early out of bag option which tests the out-of-bag samples after every iteration or the early holdout option that holds out a portion of the dataset to be used for testing at the end of every iteration	Early out of bag	early_out_of_bag, early_holdout
Learning rate	A.k.a. the gradient step, controls how aggressively the boosting algorithm fits the data. You can set values greater than 0 and smaller than 1. Large values will prevent overfitting, but smaller values generally work better (usually 10% or lower)	10%	learning_rate

Tree sampling

Option	Description	Default	API Name
Ensemble rate	The percentage of instances to build each tree in the ensemble	100%	ensemble_sample: rate
Ensemble sampling	Allows you to choose between a random sampling or a deterministic sampling to build each tree. When using deterministic sampling the random-number generator will always use the same seed, producing repeatable results.	Random	ensemble_sample: seed
Ensemble replacement	Allows a single instance to be selected multiple times to build each tree. Sampling without replacement ensures that each instance cannot be selected more than once.	True	ensemble_sample: replacement

Weighting 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 ensembles.	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 ensembles.	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 ensembles.	False	objective_weights

Dataset sampling

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 the dataset 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 an ensemble 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 ensemble. If you know that your instances are already in random order, set the shuffling to linear so that the ensemble will be constructed faster.	False	ordering:1
Random shuffling	Takes a different sampling each time you build your ensemble.	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

Operating kind options

Option	Description	Default	API Name
Probability	Averages the per-class probability distributions for all trees in the ensemble and predicts the class with higher probability. For regression ensembles, the global prediction is the mean of the individual predictions.	True	operating_kind: probability
Confidence	Averages the per-class confidences distributions for all trees in the ensemble and predicts the class with higher confidence. For regression ensembles, the global prediction is the mean of the individual predictions weighted by the expected error.	False	operating_kind: confidence
Votes	Gives one vote to each model in the ensemble. For classification models, the category with the majority of votes wins. For regression models, the global prediction is the mean of the individual predictions.	False	operating_kind: votes

Confidence, probability or votes thresholds

Option	Description	Default	API Name
Confidence, probability or votes threshold	A percentage between 0% and 100% that can be used with classification ensembles so that they only return the positive class when the confidence, probability or votes 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 for the objective field. By default BigML takes the name of the ensemble's objective field.	Objective Field Name	prediction_name
Include confidence, probability or votes	Includes an additional column with the confidence (or expected error), probability or votes in the case of Decision Forests.	False	confidence, probability, vote_count
Confidence, probability or votes column name	Sets the name you want for the confidence or expected error column. By default it is named "confidence", "probability" or "votes".	Confidence	confidence_name, probability_name, votes_name
Single tree predictions	Defines whether to include a column for each of the individual model predictions of the ensemble. This will add a column per model, named <prediction_name>_n where n is the position of the model in the model list in the ensemble, starting at 1. Only available for Decision Forests.	False	votes
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
Votes	Includes a column for each of the objective field classes indicating their votes per instance predicted. This will add a column per field, named "<objective_field_class> votes".	False	vote_counts
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 ensemble 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