BigML Dataset Cheat Sheet

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 sampling, 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

Pre-Defined Operations to Filter Datasets

Equality

Operation	Description	Field Type	API Name
Equals	Includes instances containing the specified value/values.	All	json_filter or lisp_filter
Does not equal	Excludes instances containing the specified value/values.	All	json_filter or lisp_filter

Missing

Operation	Description	Field Type	API Name
If value is missing	Includes instances containing missing values for the selected field.	All	json_filter or lisp_filter
If value is not missing	Excludes instances containing missing values for the selected field.	All	json_filter or lisp_filter

Comparison

Operation	Description	Field Type	API Name
Is between	Includes instances containing values within the specified range.	Numeric	json_filter or lisp_filter
Is less than	Includes instances containing values below the specified level.	Numeric	json_filter or lisp_filter
Is less than or equal to	Includes instances containing values equal or below the specified level.	Numeric	json_filter or lisp_filter
Is greater than	Includes instances containing values above the specified level.	Numeric	json_filter or lisp_filter
Is greater than or equal to	Includes instances containing values equal or above the specified level.	Numeric	json_filter or lisp_filter
Is between percentiles	Includes instances within the specified percentiles, e.g., a percentile between 0 and 0.3 includes the first 30% of the instances.	Numeric	json_filter or lisp_filter
Is below the mean	Includes instances below the mean of the selected field.	Numeric	json_filter or lisp_filter
Is above the mean	Includes instances above the mean of the selected field.	Numeric	json_filter or lisp_filter

Contains

Operation	Description	Field Type	API Name
Is like (case sensitive)	Matches words containing at least part of the letters specified taking into account lower and upper cases, e.g., “great” will also match text containing the word “great” or “greatness”, but not “Great” or “Greatness”.	Text	json_filter or lisp_filter
Is like (case insensitive)	Matches words containing at least part of the letters specified not taking into account lower and upper cases. e.g., “great” will also match text containing the words “great”, “greatness”, “Great” or “Greatness”.	Text	json_filter or lisp_filter
Contains (case sensitive)	Matches texts containing the exact words specified taking into account lower and upper cases. e.g., “great” will match text containing the word “great”, but not “Great”.	Text	json_filter or lisp_filter
Contains (case insensitive)	Matches texts containing the exact words specified not taking into account lower and upper cases. e.g., “great” will match text containing the word “great” or “Great”.	Text, items	json_filter or lisp_filter

Does not Contain

Operation	Description	Field Type	API Name
Is not like (case sensitive)	Checks whether given words do not match a regular expression, taking into account lower and upper cases. It is a combination of “not” and “matches?”, e.g., it will filter instances that do not contain “great” or “greatness”, but will not consider “Great”or “Greatness”.	Text	json_filter or lisp_filter
Is not like (case insensitive)	Same behavior as above but not taking into account lower and upper cases, e.g., it will filter instances that do not contain “great”, “greatness”, “Great” or “Greatness”.	Text	json_filter or lisp_filter
Not contains (case sensitive)	Checks whether a given text does not match a regular expression, taking into account lower and upper cases. It is a combination of “not” and “matches?”, e.g., it will filter instances that do not contain “great” or “greatness”, but will not consider “Great”or “Greatness”..	Text	json_filter or lisp_filter
Not contains (case insensitive)	Same behavior as above but not taking into account lower and upper cases, e.g., it will filter instances that do not contain “great”, “greatness”, “Great” or “Greatness”.	Text, items	json_filter or lisp_filter

Flatline Formula

Operation	Description	Field Type	API Name
JSON flatline formula	Computes any operation using the **Faltline** JSON syntax.	All	json_filter | field | fields
Lisp flatline formula	Computes any operation using the **Flatline** Lisp-like syntax.	All	lisp_filter | field | fields

Pre-Defined Operations to Add New Fields to Your Dataset

Discretization

Operation	Description	Field Type	API Name
Discretize by percentiles	Splits your field values into equal population segments (cate- gories). Discretizing by percentiles will split your field values into 100 different categories, by quartiles into 4, by terciles into 3, etc.	Numeric	new_fields (+ *Flatline*)
Discretize by groups	Splits the field values into equal width segments (categories), e.g., setting 3 groups for a field ranging from 0 to 6 will yield: category 1= [0, 2], category 2 = [2,4], category 3 = [4,6].	Numeric	new_fields (+ *Flatline*)
Is within percentiles?	Computes a boolean field with True or False values for each instance when you specify a percentile range between 0 and 1, depending whether they belong to the specified range or not.	Numeric	new_fields (+ *Flatline*)

Replacing Missing Values

Operation	Description	Field Type	API Name
Fixed value	Replaces all your field missing values by the specified value. You can set a number or a string.	Numeric, categorical	new_fields (+ *Flatline*)
Maximum	Replaces missing values by the max. value of the selected field.	Numeric	new_fields (+ *Flatline*)
Mean	Replaces missing values by the mean of the selected field.	Numeric	new_fields (+ *Flatline*)
Median	Replaces missing values by the median of the selected field.	Numeric	new_fields (+ *Flatline*)
Minimum	Replaces missing values by the min. value of the selected field.	Numeric	new_fields (+ *Flatline*)
Population	Replaces missing values by the number of the total instances that have valid values for the selected field, e.g., for a field containing 54 instances with valid values, the missing values will be replaced by 54.	Numeric	new_fields (+ *Flatline*)
Random integer	Replaces missing values by a random value. You can set the max. value you want for your random value generator.	Numeric	new_fields (+ *Flatline*)
Random value	Replaces missing values by a random value within your field range.	Numeric, categorical	new_fields (+ *Flatline*)
Random weighted value	Replaces missing values by a random value within the field's range, using the same probability distribution as the values in the dataset (which is described by the field's histogram).	Numeric, categorical	new_fields (+ *Flatline*)

Normalization

Operation	Description	Field Type	API Name
Normalize	Normalizes your field. Select the range for which you want to normalize your field. (This range should be within the field range.)	Numeric	new_fields (+ *Flatline*)
Z-score	Indicates the distance of the values from the mean.	Numeric	new_fields (+ *Flatline*)

Maths

Operation	Description	Field Type	API Name
Exponentiation	Computes e elevated to the field value: e^x	Numeric	new_fields (+ *Flatline*)
log_2	Scales fields logarithmically, with a logarithm base of 2. This is useful for fields with a wide range of data (since it reduces the range into a more manageable scale) and to find exponential patterns in your data.	Numeric	new_fields (+ *Flatline*)
log	Scales fields logarithmically, with a logarithm base of 10. This is useful for fields with a wide range of data (since it reduces the range into a more manageable scale) and to find exponential patterns in your data.	Numeric	new_fields (+ *Flatline*)
ln	Scales fields logarithmically, with a logarithm base of e. This is useful for fields with a wide range of data (since it reduces the range into a more manageable scale) and to find exponential patterns in your data.	Numeric	new_fields (+ *Flatline*)
Square	Squares field values: x^2	Numeric	new_fields (+ *Flatline*)
Square root	Computes the square root of the value:$\sqrt{x}$	Numeric	new_fields (+ *Flatline*)

Sliding Windos

Operation	Description	Field Type	API Name
Sum of instances	Sums consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Mean of instances	Calculates the mean of consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Median of instances	Calculates the median of consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Minimum of instances	Calculates the minimum of consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Maximum of instances	Calculates the maximum of consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Product of instances	Calculates the product of consecutive instances by defining a window start and end (negative values are previous instances and positive values next instances).	Numeric	new_fields (+ *Flatline*)
Difference from first	Calculates the difference between values associated with the start and end indices of the window, where the end index must be greater than the start index and the difference is calculated as end - start.	Numeric	new_fields (+ *Flatline*)
Difference from first (%)	Calculates the percentage difference between values associated with the start and end indices of the window, where the end index must be greater than the start index and the difference is calculated as end - start.	Numeric	new_fields (+ *Flatline*)
Difference from last	Calculates the difference between values associated with the start and end indices of the window, where the end index must be greater than the start index and the difference is calculated as start - end.	Numeric	new_fields (+ *Flatline*)
Difference from last (%)	Calculates the percentage difference between values associated with the start and end indices of the window, where the end index must be greater than the start index and the difference is calculated as start - end.	Numeric	new_fields (+ *Flatline*)

Types

Operation	Description	Field Type	API Name
Categorical	Coerces numeric field values into categorical values, e.g., the number 10 will become a string "10".	Numeric	new_fields (+ *Flatline*)
Integer	Coerces categorical values to integer values, e.g., the string "7.5 pounds" will become 7. Boolean values are assigned 0 (false) and 1 (true).	Categorical, text, items	new_fields (+ *Flatline*)
Real	Coerces categorical values to floating point values, e.g., the string "7.5 pounds" will become 7.5. Boolean values are assigned 0 and 1.	Categorical, text, items	new_fields (+ *Flatline*)

Random

Operation	Description	Field Type	API Name
Random integer	Computes a random integer for each instance.	Numeric	new_fields (+ *Flatline*)
Random value within field range	Computes a random value by taking your field range as the reference for min. and max. values.	Numeric, categorical	new_fields (+ *Flatline*)
Random weighted value	Computes a random value within the field's range, using the same probability distribution as the values in the dataset (which is described by the field's histogram).	Numeric, categorical	new_fields (+ *Flatline*)

Statistics

Operation	Description	Field Type	API Name
Mean	Computes the field mean for all instances.	Numeric	new_fields (+ *Flatline*)
Population	Computes the count of total instances for that field.	Numeric	new_fields (+ *Flatline*)
Population fraction	Computes the number of instances with values below the specified value.	Numeric	new_fields (+ *Flatline*)

Flatline Formula

Operation	Description	Field Type	API Name
JSON flatline formula	Computes any operation using the **Faltline** JSON syntax.	All	new_fields["field"]
Lisp flatline formula	Computes any operation using the **Flatline** Lisp-like syntax.	All	new_fields["field"]

Remove Duplicates

Option	Description	Field Type	API Name
Remove duplicates	Removes the duplicated instances in your dataset taking into account all the field values.	All	*SQL query*

Aggregate Instances

Option	Description	Field Type	API Name
Agreggating field	Sets the field of the dataset that you want to use to group your instances.	All	*SQL query*
Operation	Specifies the aggregation operation to be applied to the rest of the dataset fields.	All	*SQL query*

Join Datasets

Option	Description	Field Type	API Name
Type of join	Allows you to select a left join, right join, full join, and inner join.	N/A	*SQL query*
Selected dataset	Allows you to select the dataset you want to join with the current dataset.	N/A	*SQL query*
Join fields	Allows you to select one or more fields from the current dataset to match the instances with the selected dataset. These fields should have the same values in both datasets so the instances can be matched.	All	*SQL query*
Select fields	Allows you to choose all the fields from the selected dataset or select a subset of them.	All	*SQL query*
Filter datasets	Allows you to filter the current and/or selected dataset while making the join.	All	*SQL query*

Merge Datasets

Option	Description	Default	API Name
Merge datasets	Select up to 32 datasets with the same fields to merge their instances into one dataset.	N/A	*SQL query*
Rate	Sets the proportion of each merging dataset you want to consider between 0% and 100%.	100%	sample_rate
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 out of bag. It is only selectable when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag
Seed	Allows you to set a string to produce repeatable results.	None	seed

Order Instances

Option	Description	Default	API Name
Order instances	Allows you to sort the rows of a dataset by one or more selected fields in ascending or descending order	N/A	*SQL query*