# H2O AutoML

AutoML is a function in H2O that automates the process of building a large number of models, with the goal of finding the "best" model without any prior knowledge or effort by the Data Scientist.

The H2O AutoML interface is designed to have as few parameters as possible so that all the user needs to do is point to their dataset, identify the response column and optionally specify a time constraint or limit on the number of total models trained.

In both the R and Python API, AutoML uses the same data-related arguments, x, y, training\_frame, validation\_frame, as the other H2O algorithms. Most of the time, all you’ll need to do is specify the data arguments. You can then configure values for max\_runtime\_secs and/or max\_models to set explicit time or number-of-model limits on your run.

**Required Data Parameters**

* [y](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/y.html): This argument is the name (or index) of the response column.
* [training\_frame](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/training_frame.html): Specifies the training set.

**Required Stopping Parameters**

One of the following stopping strategies (time or number-of-model based) must be specified. When both options are set, then the AutoML run will stop as soon as it hits one of either of these limits.

* [max\_runtime\_secs](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/max_runtime_secs.html): This argument controls how long the AutoML will run at the most, before training the final Stacked Ensemble models. Defaults to 3600 seconds (1 hour).
* [max\_models](http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/max_models.html): Specify the maximum number of models to build in an AutoML run, excluding the Stacked Ensemble models. Defaults to NULL/None.

## AutoML Output

The AutoML object includes a “leaderboard” of models that were trained in the process, including the 5-fold cross-validated model performance (by default). The number of folds used in the model evaluation process can be adjusted using the nfolds parameter. If the user would like to score the models on a specific dataset, they can specify the leaderboard\_frame argument, and then the leaderboard will show scores on that dataset instead.

The models are ranked by a default metric based on the problem type (the second column of the leaderboard). In binary classification problems, that metric is AUC, and in multiclass classification problems, the metric is mean per-class error. In regression problems, the default sort metric is deviance. Some additional metrics are also provided, for convenience.

# Our take on AutoML

AutoML is here to stay. I am eager to see the direction where it goes to further advancements in data science. A single automated mixer certainly cannot outperform a human creative mind when it comes to feature engineering but in my experience, AutoML is worth exploring.

Although AutoML alone won’t get you top spot in machine learning competitions, it is definitely worth considering as an addition alongside your blended and stacked models. In recent competitions, the AutoML model boosted my score considerably which led me to explore and concentrate on the blending part.