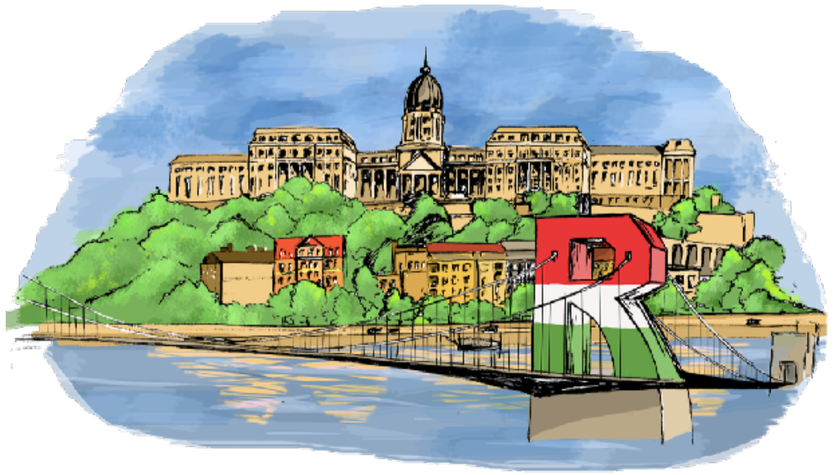
Scalable Automatic Machine Learning in H2O



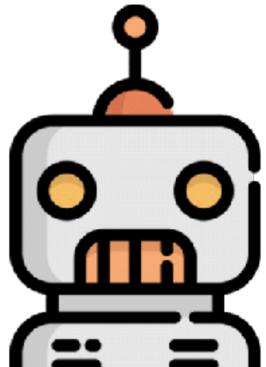
2018 European R Users Meeting
Budapest, Hungary

Erin LeDell Ph.D.
@ledell

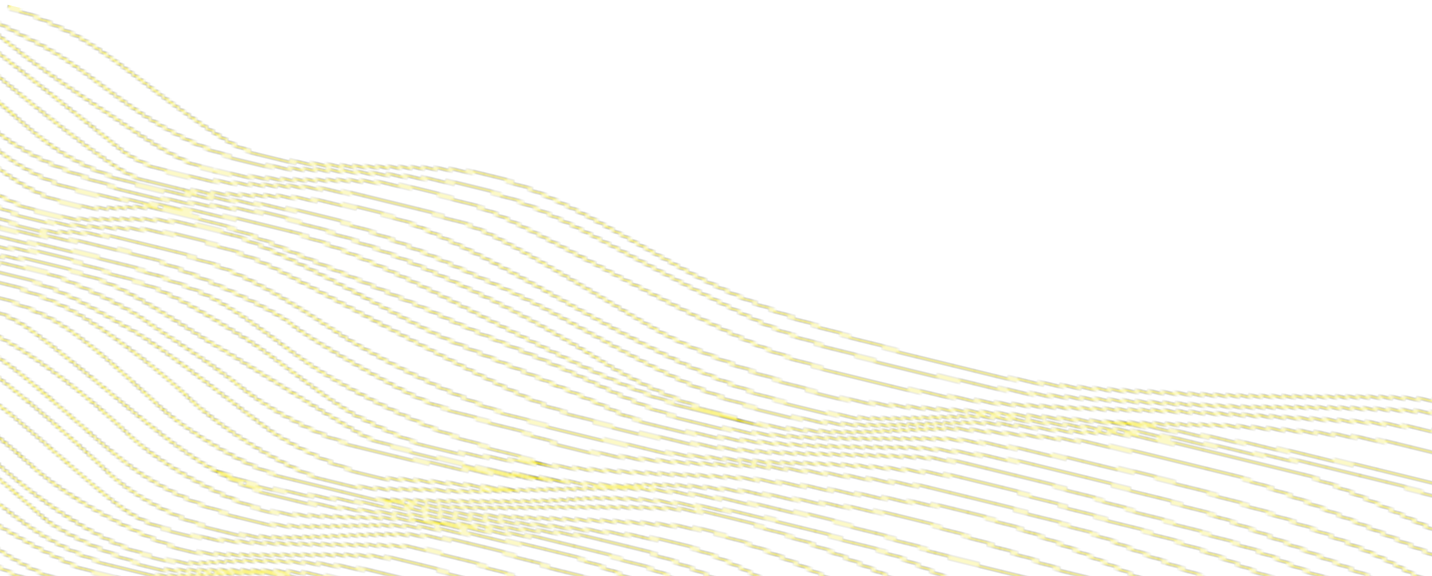
Agenda

- Intro to Automatic Machine Learning (AutoML)
- H2O AutoML Overview
- Pro Tips
- Tutorials

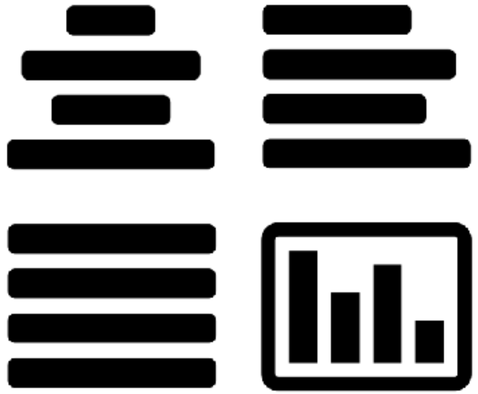
Slides ⬇️ https://tinyurl.com/aml-erum



Intro to Automatic Machine Learning

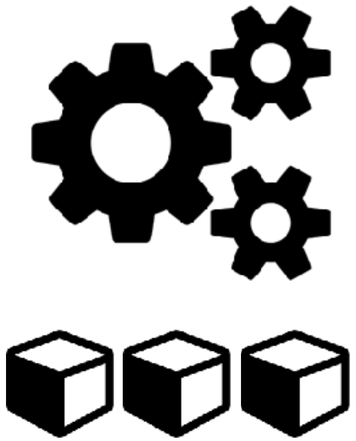


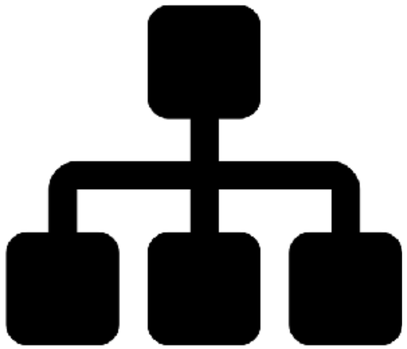
Aspects of Automatic Machine Learning



Data Prep

Model Generation





Ensembles

Aspects of Automatic Machine Learning

Data Preprocessing

- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features

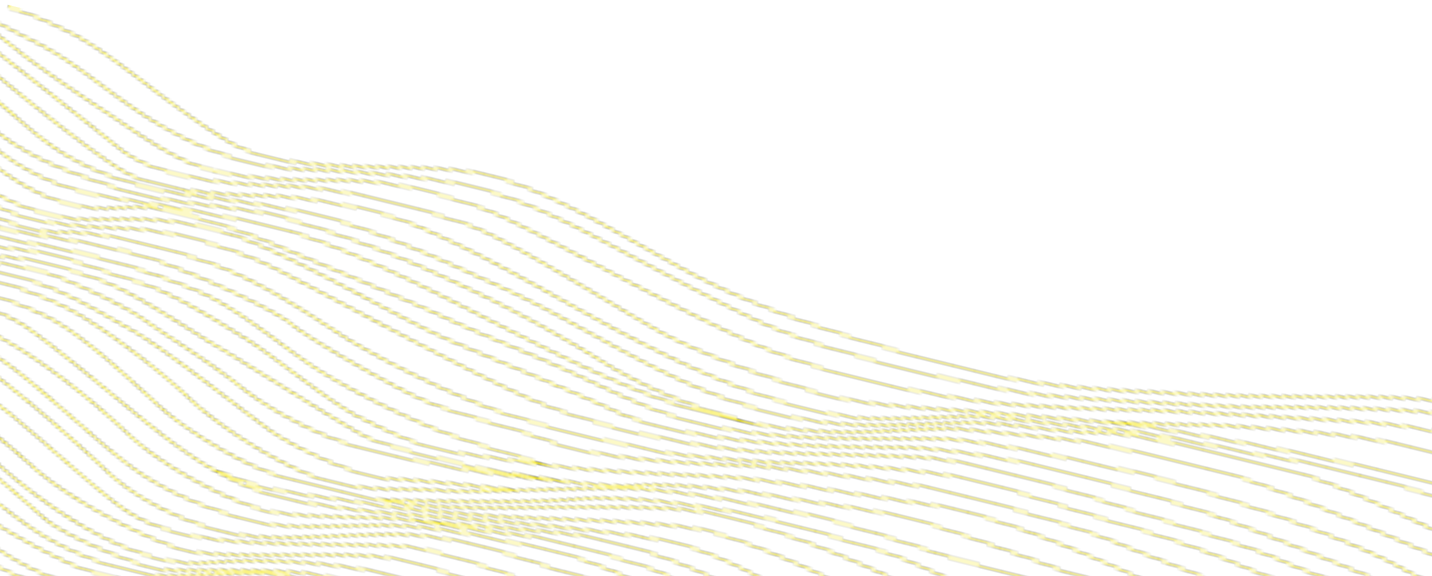
Model Generation

- Cartesian grid search or random grid search
- Bayesian Hyperparameter Optimization
- Individual models can be tuned using a validation set

Ensembles

- Ensembles often out-perform individual models
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

H2O's AutoML



H2O Machine Learning Platform

- Open source, distributed (multi-core + multi-node) implementations of cutting edge ML algorithms.
- Core algorithms written in high performance Java.
- APIs available in R, Python, Scala; web GUI.
- Easily deploy models to production as pure Java code.
- Works on Hadoop, Spark, AWS, your laptop, etc.

H2O AutoML (current release v3.18)

Data Preprocessing

- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features

Model Generation

- Cartesian grid search or random grid search
- Bayesian Hyperparameter Optimization
- Individual models can be tuned using a validation set

Ensembles

- Ensembles often out-perform individual models:
- Stacking / Super Learning (Wolpert, Breiman)
- Ensemble Selection (Caruana)

Random Grid Search & Stacking

- Random Grid Search combined with Stacked Ensembles is a powerful combination.
- Ensembles perform particularly well if the models they are based on (1) are individually strong, and (2) make uncorrelated errors.
- Stacking uses a second-level metalearning algorithm to find the optimal combination of base learners.

H2O AutoML

- Basic data pre-processing (as in all H2O algos).
- Trains a random grid of GBMs, DNNs, GLMs, etc. using a carefully chosen hyper-parameter space
- Individual models are tuned using a validation set.
- Two Stacked Ensembles are trained ("All Models" ensemble & a lightweight "Best of Family" ensemble).
- Returns a sorted "Leaderboard" of all models.

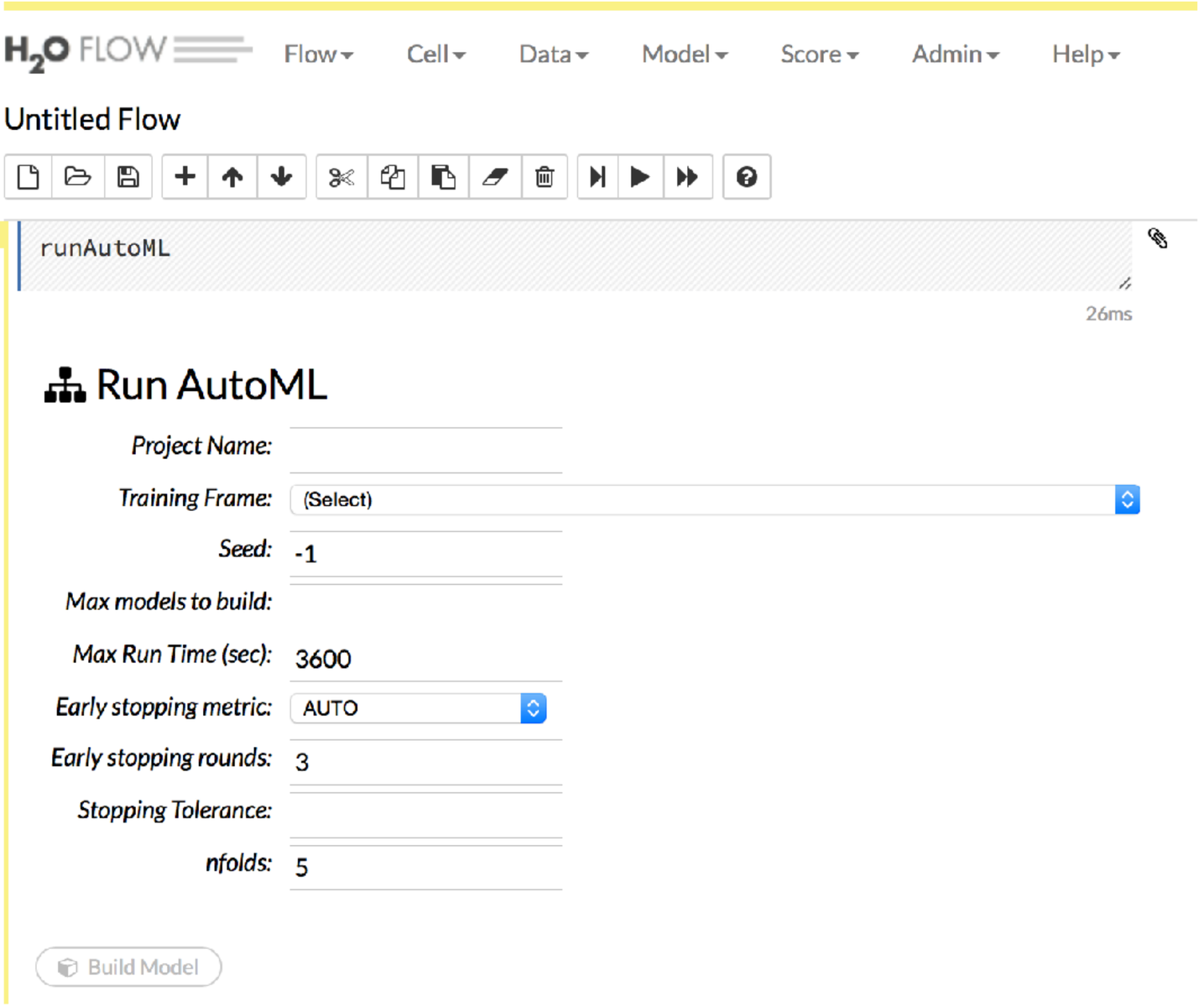
Available in H2O >=3.14

H2O AutoML in R

Example

```
library(h2o)

h2o.init()

train <- h2o.importFile("train.csv")


aml <- h2o.automl(y = "response_colname",
                  training_frame = train,
                  max_runtime_secs = 600)

lb <- aml@leaderboard
```

H2O AutoML in Flow GUI

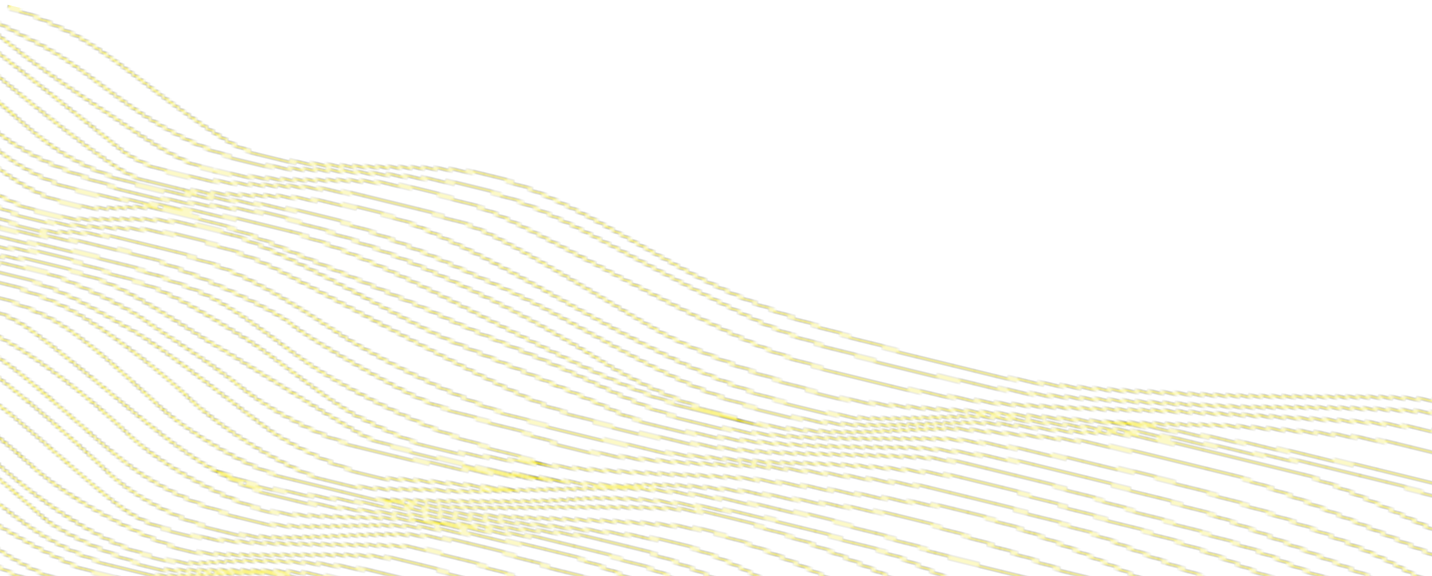


H2O AutoML Leaderboard

model_id	auc	logloss
StackedEnsemble_AllModels_0_AutoML_20171121_012135	0.788321	0.554019
StackedEnsemble_BestOfFamily_0_AutoML_20171121_012135	0.783099	0.559286
GBM_grid_0_AutoML_20171121_012135_model_1	0.780554	0.560248
GBM_grid_0_AutoML_20171121_012135_model_0	0.779713	0.562142
GBM_grid_0_AutoML_20171121_012135_model_2	0.776206	0.564970
GBM_grid_0_AutoML_20171121_012135_model_3	0.771026	0.570270
DRF_0_AutoML_20171121_012135	0.734653	0.601520
XRT_0_AutoML_20171121_012135	0.730457	0.611706
GBM_grid_0_AutoML_20171121_012135_model_4	0.727098	0.666513
GLM_grid_0_AutoML_20171121_012135_model_0	0.685211	0.635138

Example Leaderboard for binary classification

H2O AutoML Pro Tips!



Before you press the “red button”

AutoML Pro Tips: Input Data

- Don't use `leaderboard_frame` unless you really need to; use cross-validation metrics to generate the leaderboard instead (default).
- If you only provide `training_frame`, it will chop off 20% of your data for a validation set to be used in early stopping. To control this proportion, you can split the data yourself and pass a `validation_frame` manually.

AutoML Pro Tips: Exclude Algos

- If you have sparse, wide data (e.g. text), use the `exclude_algos` argument to turn off the tree-based models (GBM, RF).
- If you want tree-based algos only, turn off GLM and DNNs via `exclude_algos`.

AutoML Pro Tips: Time & Model Limits

- AutoML will stop after 1 hour unless you change `max_runtime_secs`.
- Running with `max_runtime_secs` is not reproducible since available resources on a machine may change from run to run. Set `max_runtime_secs` to a big number (e.g. 999999999) and use `max_models` instead.

AutoML Pro Tips: H2O cluster memory

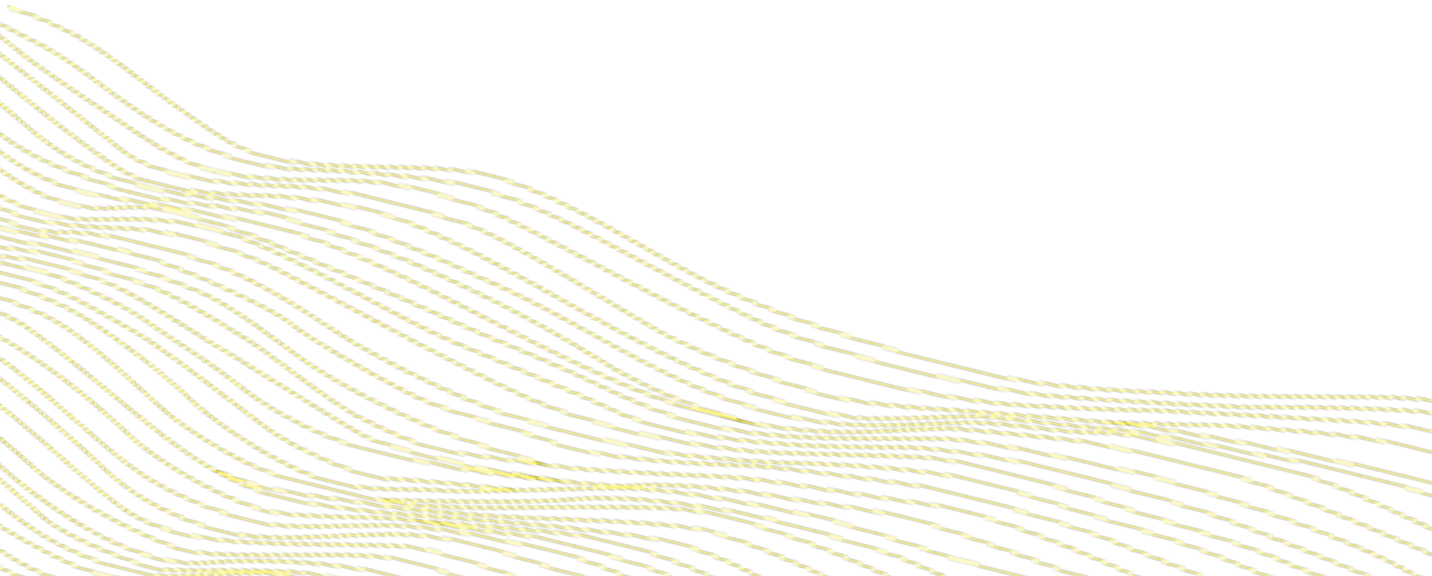
- Reminder: All H2O models are stored in H2O Cluster memory.
- Make sure to give the H2O Cluster a lot of memory if you're going to create hundreds or thousands of models.
- e.g. `h2o.init(max_mem_size = "80G")`

AutoML Pro Tips: Saving Models

- You can save any of the individual models created by the AutoML run. The model ids are listed in the leaderboard.

- If you're taking your leader model (probably a Stacked Ensemble) to production, we'd recommend using "Best of Family" since it only contains 5 models and gets most of the performance of the "All Models" ensemble.

H2O AutoML Tutorials



H2O AutoML Tutorials

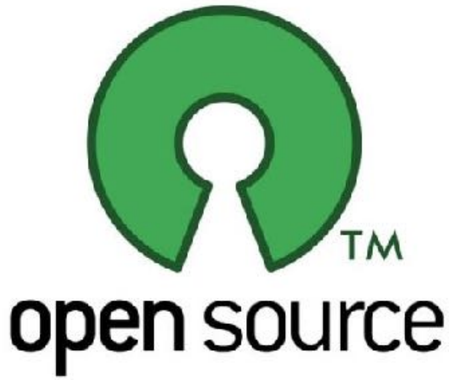
https://tinyurl.com/h2o-automl-tutorials

Code available here

H2O Resources

- Documentation: http://docs.h2o.ai
- Tutorials: https://github.com/h2oai/h2o-tutorials
- Slidedecks: https://github.com/h2oai/h2o-meetups
- Videos: https://www.youtube.com/user/0xdata
- Stack Overflow: https://stackoverflow.com/tags/h2o
- Google Group: https://tinyurl.com/h2ostream
- Gitter: http://gitter.im/h2oai/h2o-3
- Events & Meetups: http://h2o.ai/events

Contribute to H2O!



Get in touch over email, Gitter or JIRA.

https://tinyurl.com/h2o-contribute

Thank you!

@ledell on Github, Twitter

erin@h2o.ai

http://www.stat.berkeley.edu/~ledell