

# Scalable Automatic Machine Learning with H2O AutoML

Erin LeDell, Ph.D.  
Chief Machine Learning Scientist  
H2O.ai  
@ledell



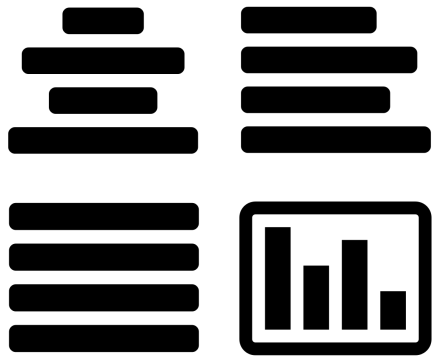
# Agenda

- Intro to Automatic Machine Learning (AutoML)
- H2O AutoML Overview
- AutoML Pro Tips
- Q & A

Slides  <https://tinyurl.com/automl-nyc>

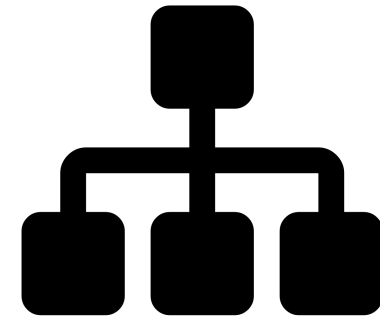
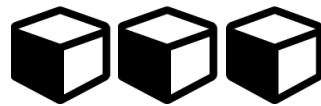
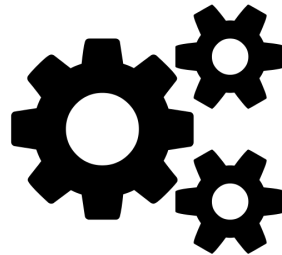


# Aspects of Automatic Machine Learning



Data Prep

Modeling



Ensembles

# Aspects of Automatic Machine Learning

## Data Prep

- Imputation, one-hot encoding, normalization
  - Feature selection, dimensionality reduction
  - Count/Label/Target encoding of categorical features
- 

## Modeling

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

## Ensembles

- Stacking / Super Learning (Wolpert, Breiman, van der Laan)
- Ensemble Selection (Caruana)
- Blending



# 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 features)

## Data Prep

- Imputation, one-hot encoding, normalization
  - Feature selection, dimensionality reduction
  - Count/Label/Target encoding of categorical features
- 

## Modeling

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

## Ensembles

- Stacking / Super Learning (Wolpert, Breiman, van der Laan)
- Ensemble Selection (Caruana)
- Blending



# H2O AutoML

- Basic data pre-processing (as in all H2O algos).
- Trains a random grid of GBMs and DNNs, using a carefully chosen hyper-parameter space. Also tunes a GLM, and includes a Random Forest and a Random Forest with “Extremely Randomized Trees”.
- Individual models are tuned using a validation set.
- Two Stacked Ensembles are trained (“All Models” ensemble and a lightweight “Best of Family” ensemble).
- Returns a sorted “Leaderboard” of all models.

AutoML available in H2O  $\geq 3.14$



# H2O AutoML Roadmap


- Addition of XGBoost.
- Basic target encoding of high cardinality categorical features.
- Improvements to the random grid search hyperspace.
- Improvements for imbalanced datasets.
- New ensembles.
- Public benchmarks.

Coming soon!






# H2O AutoML Flow GUI

**H<sub>2</sub>O FLOW**  Flow ▾ Cell ▾ Data ▾ Model ▾ Score ▾ Admin ▾ Help ▾


Untitled Flow



CS


runAutoML

26ms

 **Run AutoML**

Project Name:

Training Frame:


(Select) 

Seed:

Max models to build:

Max Run Time (sec):


Early stopping metric:

AUTO 

Early stopping rounds:

Stopping Tolerance:

nfolds:





# 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 Python

## Example

```
import h2o
from h2o.automl import H2OAutoML
h2o.init()

train = h2o.import_file("train.csv")

aml = H2OAutoML(max_runtime_secs = 600)
aml.train(y = "response_colname",
          training_frame = train)

lb = aml.leaderboard
```



# 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



# Before you press “Go”: AutoML Pro Tips



# H2O AutoML Pro Tips

- 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.
- Don't use `leaderboard_frame` unless you really need to; use cross-validation metrics to generate the leaderboard instead (default). This is the most efficient use of your data.



# H2O AutoML Pro Tips

- Specific algorithms can be excluded from AutoML using the `exclude_algos` argument. If you have sparse, wide data (e.g. text), use the `exclude_algos` argument to turn off the tree-based models (GBM, RF), or apply Word2Vec or PCA first to reduce dimensionality. If you want tree-based models only, turn off GLM and DNNs.
- If you have time-series data, don't use cross-validation or Stacked Ensembles. Turn off CV by setting `nfolds = 0` (which will automatically disable Stacked Ensembles). Also remember that with time-series data, the training data should precede the validation data (in time).



# H2O AutoML Pro Tips

- AutoML will stop after 1 hour unless you change `max_runtime_secs`.
- AutoML limited only by time (with `max_runtime_secs`) is not reproducible since available resources on a machine may change from run to run.
- For reproducibility, set `max_runtime_secs` to a big number (e.g. 999999999) and limit by the number of models using `max_models` instead.





# H2O AutoML Pro Tips

- Reminder: All H2O models are stored in H2O Cluster memory. Make sure to give the H2O Cluster adequate memory if you're planning on training hundreds or thousands of models.  
e.g. `h2o.init(max_mem_size = "80G")`
- If you want to add (train) more models to an existing AutoML project, just make sure to use the same training set and `project_name` and run AutoML again. If you set the same seed twice it will give you identical models as the first run (not useful), so change the seed or leave it unset.



# H2O AutoML Pro Tips

- If you're expecting more models than are listed in the leaderboard, or the run is stopping earlier than `max_runtime_secs`, this is a result of the default “early stopping” settings.
- To allow more time, increase the number of `stopping_rounds` and/or decrease value of `stopping_tolerance`.
- The leaderboard will rank models by a default metric, however you can change the leaderboard `sort_metric` and also the `stopping_metric` to optimize for a metric of your choice.



# H2O AutoML Pro Tips

- You can save any of the individual models created by the AutoML run to disk using the standard H2O model saving functions (binary format or MOJO/POJO). The model ids are listed in the leaderboard, and all the models from the AutoML run remain accessible in the H2O Cluster.
- If you're taking your leader model (probably a Stacked Ensemble) to production, we'd recommend using "Best of Family" ensemble since it only contains 5 models and gets most of the performance of the "All Models" ensemble.



# H2O AutoML Tutorials



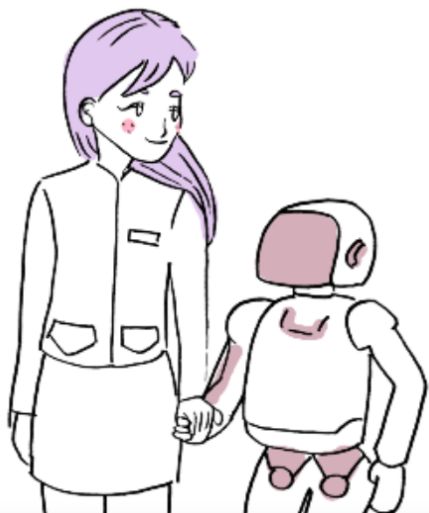
<https://tinyurl.com/automl-tutorials>



# Thank you!

@ledell  

erin@h2o.ai



Slides  <https://tinyurl.com/automl-nyc>

