# $H_2O$

# WORLD

2017

# Erin LeDell

## Machine Learning Scientist

@ledell / erin@h2o.ai

# AutoML

Automatic Machine Learning

# MEET THE MAKERS

ERIN LEDELL

Machine Learning Scientist

NAVDEEP GILL

Software Engineer
& Data Scientist

RAY PECK

Director of Product
Engineering

H$_2$O WORLD 2017

# Agenda

- Intro to Automatic Machine Learning (AutoML)
- Random Grid Search & Stacked Ensembles
- H2O's AutoML (R, Python, GUI)
- H2O-3 Roadmap
- Hands-on Tutorial

# Intro to AutoML

## Automatic Machine Learning

# Aspects of Automatic ML

Data Prep

Model Generation

Ensembles

# Data Prep

- Imputation of missing data
- Standardization of numeric features
- One-hot encoding of categorical features
- Count/Label/Target encoding of categorical features
- Feature selection and/or feature extraction (e.g. PCA)
- Feature engineering

# Model Generation

- Cartesian grid search
- Random grid search
- Tune individual models via Early Stopping
- Bayesian Hyperparameter Optimization

# Ensembles

- Bagging / Averaging
- Stacking / Super Learning
- Ensemble Selection

# Random Stacking

Random Grids + Stacked Ensembles

# Stacked Ensembles

$$n\left\{\overbrace{\begin{bmatrix} \quad & X & \quad \end{bmatrix}}^{m}\right. \begin{bmatrix} y \end{bmatrix}$$

- Specify L base learners (with model params).
- Specify a metalearner (just another algo).
- Perform k-fold cross-validation on the base learners.

# Stacked Ensembles

$$\mathrm{n}\left\{\begin{bmatrix} \\ p_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ p_L \\ \\ \end{bmatrix} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}\right. \rightarrow \mathrm{n}\left\{\overbrace{\begin{bmatrix} & & \\ & Z & \\ & & \end{bmatrix}}^{\mathrm{L}} \begin{bmatrix} \\ y \\ \\ \end{bmatrix}\right.$$

- Collect cross-validated predicted values from base learners.
- Train a second-level metalearning algorithm to find the optimal combination of base learners.
- Metalearner requires only a small amount of compute on top of the cross-validation process (it’s cheap).

# Random Grid Search + Stacking

- Random Grid Search combined with Stacked Ensembles is a powerful combination.
- Stacked Ensembles perform particularly well if the models they are based on (1) are ***individually strong***, and (2) ***make uncorrelated errors***.
- Random Grid Search is an excellent way to create a diverse of models for the ensemble.

# H2O AutoML

Automatic Machine Learning in H2O

# H2O Machine Learning Platform

- 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.
- Works on Hadoop, Spark, EC2, your laptop, etc.
- Easily deploy models to production as pure Java code.

# H2O AutoML (first cut)

- Imputation, one-hot encoding, standardization.
- Random Grid Search over a custom hyperparameter space, defined by expert data scientists.
- Early stopping of individual models and random grids.
- GBMs, Random Forests, Deep Neural Nets, GLMs
- Multiple Stacked Ensembles of models.
- Leaderboard for ranking.

# H2O AutoML in R

### Example

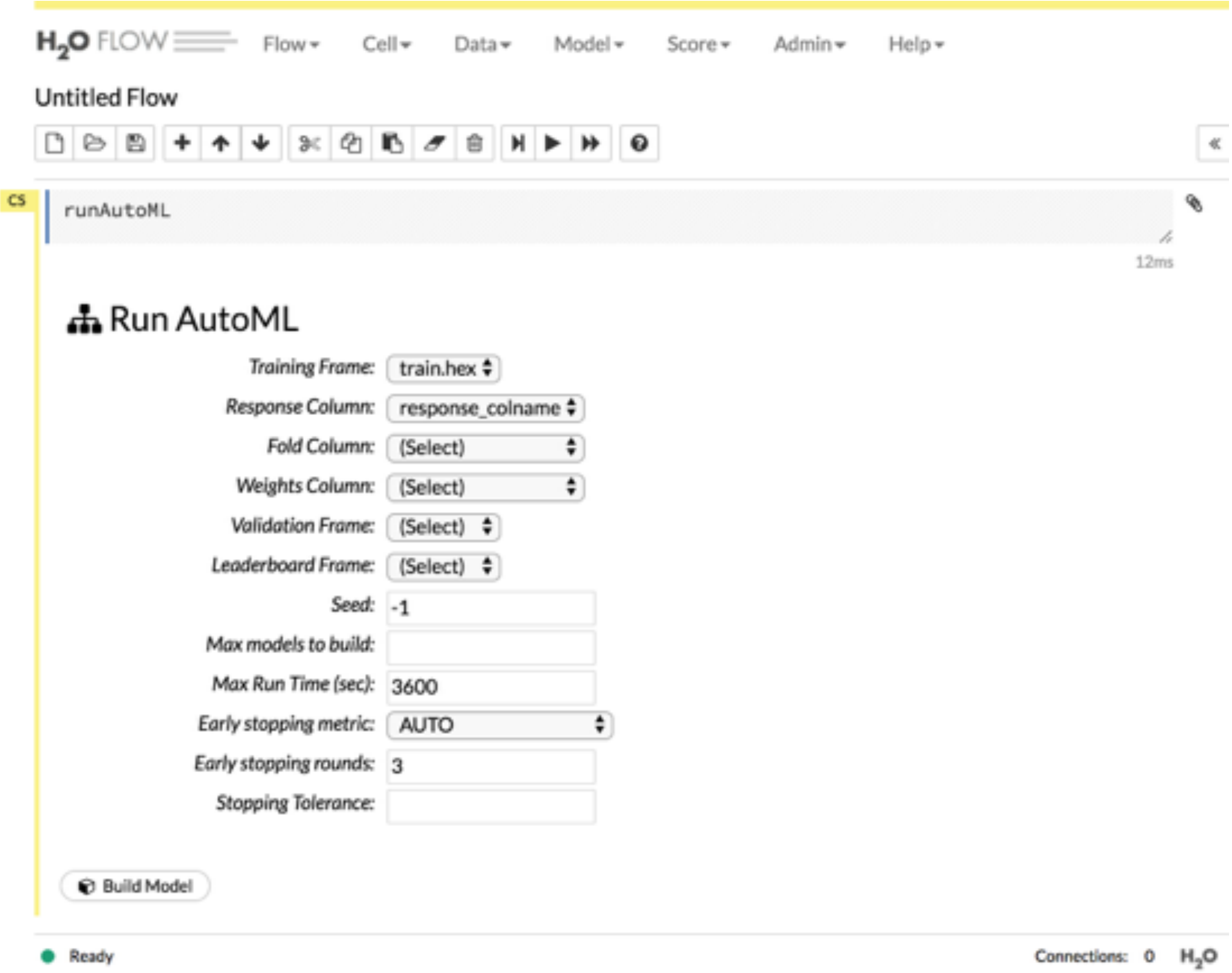
```r
library(h2o)

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

```python
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 in Flow



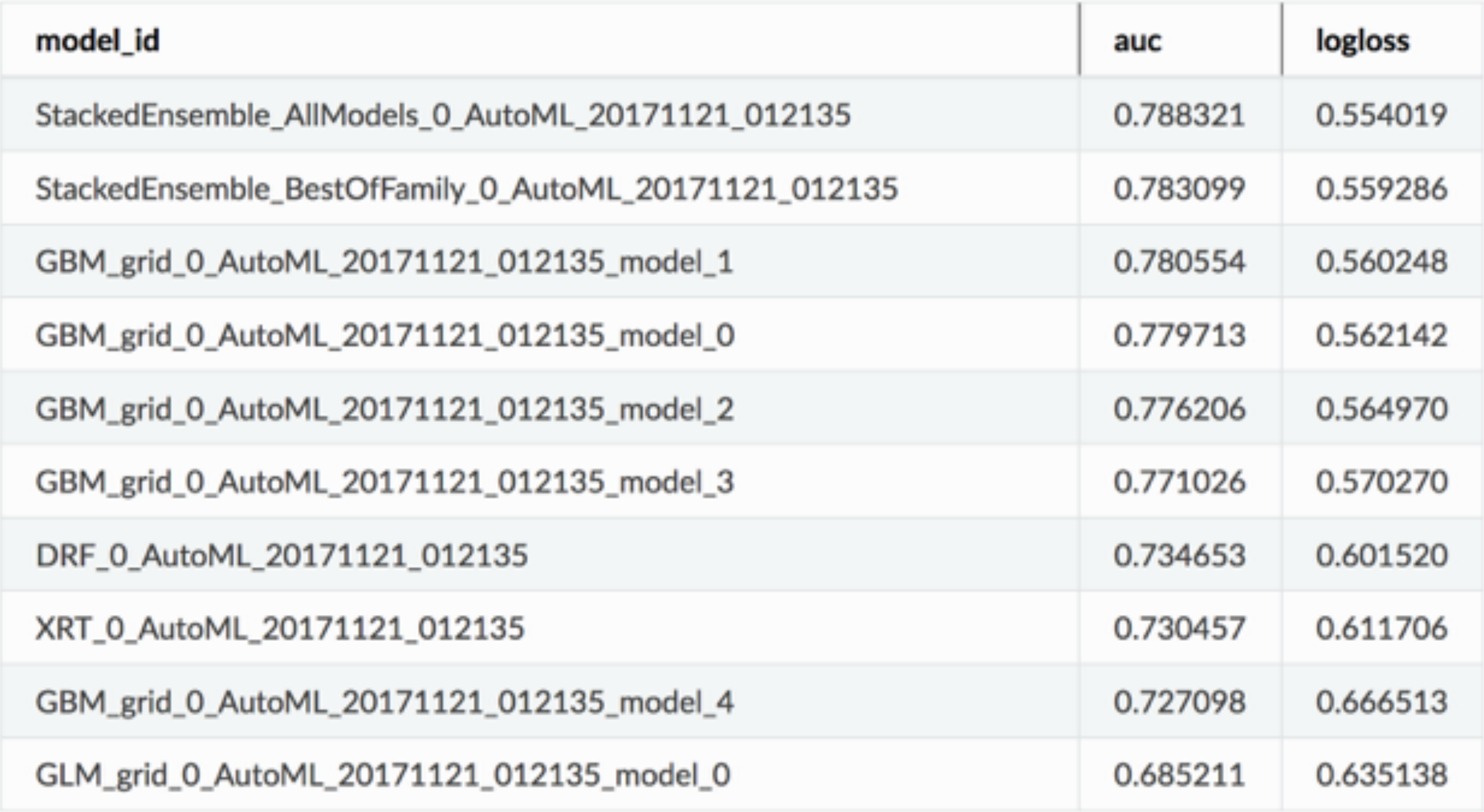
```
runAutoML
```

## Run AutoML

| | |
|---|---|
| Training Frame: | train.hex |
| Response Column: | response_colname |
| Fold Column: | (Select) |
| Weights Column: | (Select) |
| Validation Frame: | (Select) |
| Leaderboard Frame: | (Select) |
| Seed: | -1 |
| Max models to build: | |
| Max Run Time (sec): | 3600 |
| Early stopping metric: | AUTO |
| Early stopping rounds: | 3 |
| Stopping Tolerance: | |

Build Model

# 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-3 Roadmap

## Coming Soon to H2O

# H2O-3 Roadmap

| Feature | Q1 | Q2 |
|---|---|---|
| *New Algorithm: Cox-Proportional Hazards* | 🟩 | 🟩 |
| *GLM: Ordinal Regression* | 🟩 | 🟩 |
| *GBM: Quasibinomial* | 🟩 | 🟩 |
| *NLP Improvements, TF-IDF* | 🟩 | 🟩 |
| *Stacked Ensemble: Custom Metalearner* | 🟩 | 🟩 |
| *AutoML: New Ensembles* | 🟩 | 🟩 |
| *AutoML: Add XGBoost* | | 🟩 |
| *Distributed XGBoost* | | 🟩 |
| *New Algorithm: Factorization Machines* | | 🟩 |

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

- 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
- Events & Meetups: http://h2o.ai/events
- Stack Overflow: https://stackoverflow.com/tags/h2o
- Google Group: https://tinyurl.com/h2ostream
- Gitter: http://gitter.im/h2oai/h2o-3

# DEMO

Hands-on Tutorial

# First-time Qwiklab Account Setup

- Go to http://h2oai.qwiklab.com
- Click on “JOIN”
- Create a new account with a valid email address
- You will receive a confirmation email
  - Click on the link in the confirmation email
- Go back to http://h2oai.qwiklab.com and log in
- Go to the Catalog on the left bar
- Choose “Introduction to AutoML in H2O”
- Wait for instructions

# H2O AutoML Tutorial

https://tinyurl.com/automl-h2oworld17

Code and data available here