Automatic Machine Learning in R with H2O

NYR 2020

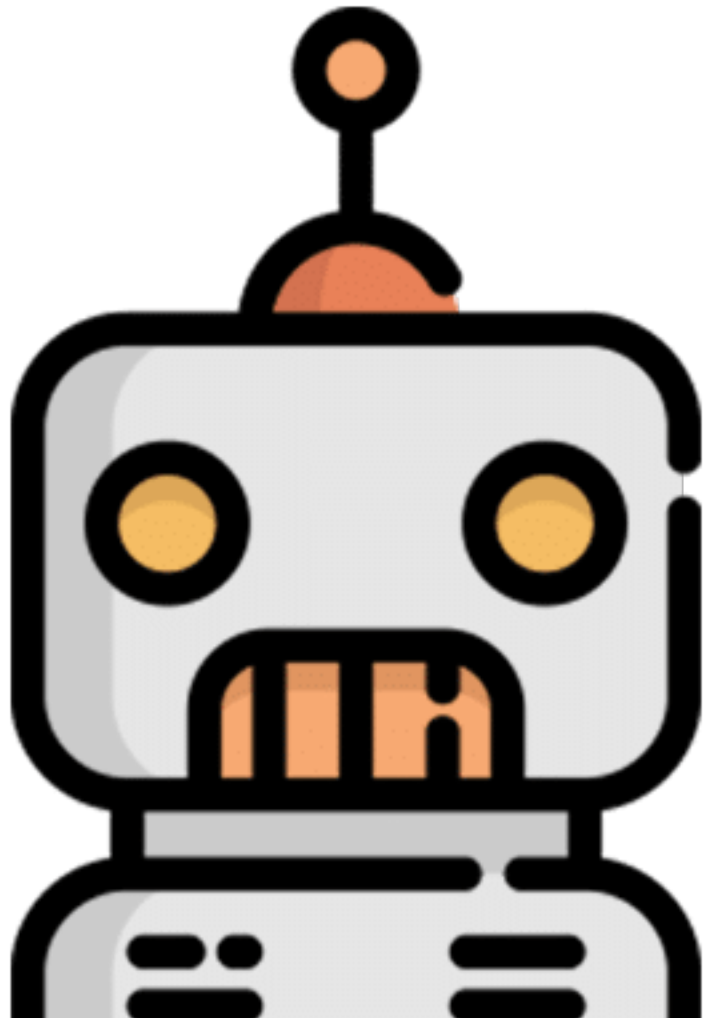
H2O.ai

Erin LeDell Ph.D.
@ledell

Agenda

- H2O Platform
- Automatic Machine Learning (AutoML)
- H2O AutoML Overview
- Resources

Slides ⬇️ https://tinyurl.com/h2o-meetups



H2O Platform

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

H2O Machine Learning Features

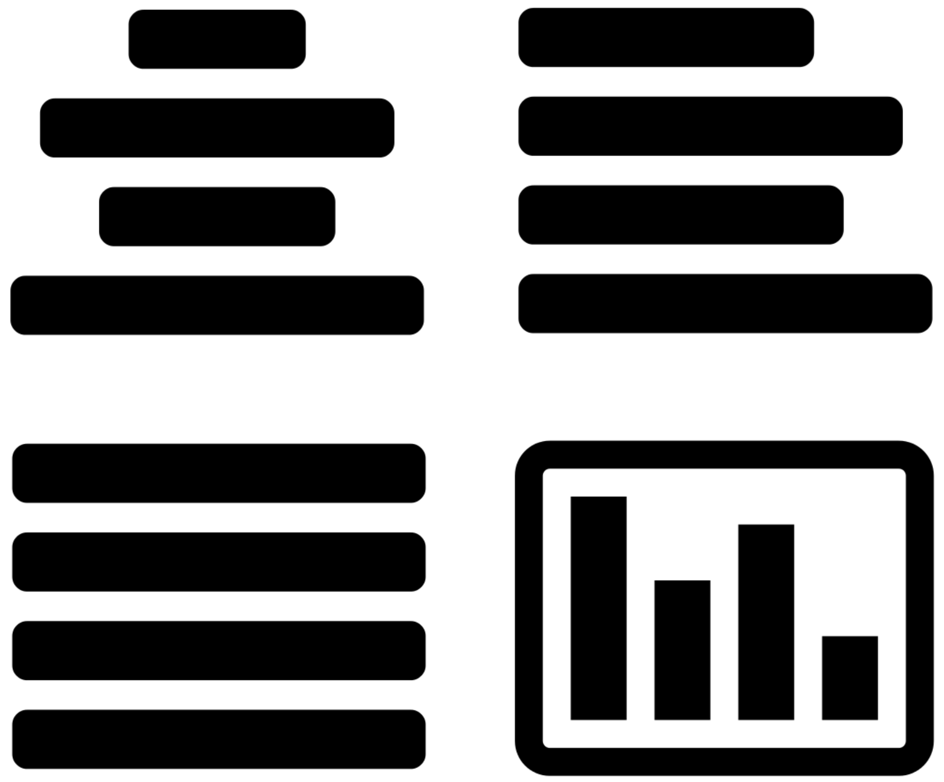
- Supervised & unsupervised machine learning algos (GBM, RF, DNN, GLM, Stacked Ensembles, etc.)
- Imputation, normalization & auto one-hot-encoding
- Automatic early stopping
- Cross-validation, grid search & random search
- Variable importance, model evaluation metrics, plots

Intro to Automatic Machine Learning

Goals & Features of AutoML

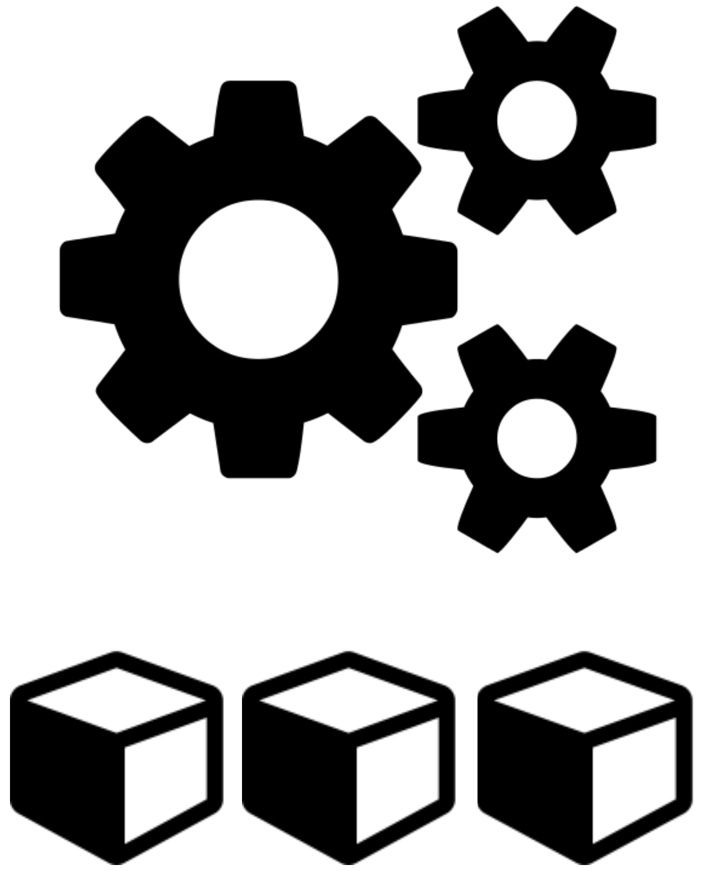
- 🏆 Train the best model in the least amount of time.
- 📉 Reduce the human effort & expertise required in machine learning.
- 📈 Improve the performance of machine learning models.
- 🔄 Increase reproducibility & establish a baseline for scientific research or applications.

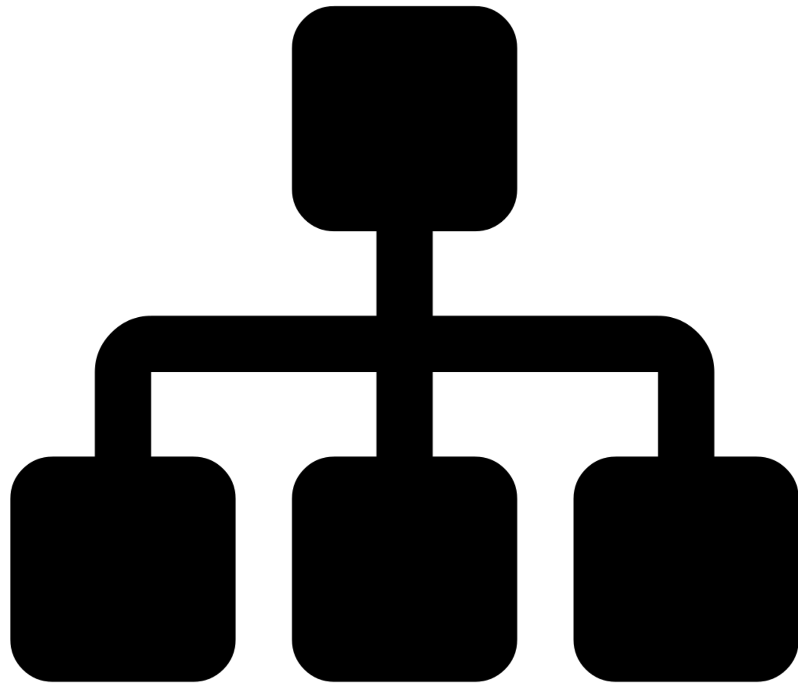
Aspects of Automatic Machine Learning



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)

Different Flavors of AutoML

August 15th, 2018

The different flavors of AutoML

Share **Category:** AutoML, Data Science, Driverless AI, H2O

By: Erin LeDell

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software (e.g. H2O, scikit-learn, keras). Although these tools have made it easy to train and evaluate machine learning models, there is still a good amount of data science knowledge that's required in order to create the *highest-quality* model, given your dataset. Writing the code to perform a hyperparameter search over many different types of algorithms can also be time consuming and repetitive work.

What is AutoML?

https://tinyurl.com/flavors-of-automl

H2O AutoML

H2O AutoML (v3.30)

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) <u>are individually strong</u>, and (2) <u>make uncorrelated errors</u>.
- 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 cross-validation.
- Two Stacked Ensembles are trained ("All Models" ensemble & a lightweight "Best of Family" ensemble).
- Returns a sorted "Leaderboard" of all models.
- All models can be easily exported to production.

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 Leaderboard

model_id	auc	logloss	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20181212_105540	0.7898014	0.5511086	0.3331737	0.4321104	0.1867194
StackedEnsemble_BestOfFamily_AutoML_20181212_105540	0.7884246	0.5521454	0.3231919	0.4326254	0.1871647
XGBoost_1_AutoML_20181212_105540	0.7846510	0.5575305	0.3254707	0.4349489	0.1891806
XGBoost_grid_1_AutoML_20181212_105540_model_4	0.7835232	0.5578542	0.3188188	0.4352486	0.1894413
XGBoost_grid_1_AutoML_20181212_105540_model_3	0.7830043	0.5596125	0.3250808	0.4357077	0.1898412
XGBoost_2_AutoML_20181212_105540	0.7813603	0.5588797	0.3470738	0.4359074	0.1900153
XGBoost_3_AutoML_20181212_105540	0.7808475	0.5595886	0.3307386	0.4361295	0.1902090
GBM_5_AutoML_20181212_105540	0.7808366	0.5599029	0.3408479	0.4361915	0.1902630
GBM_2_AutoML_20181212_105540	0.7800361	0.5598060	0.3399258	0.4364149	0.1904580
GBM_1_AutoML_20181212_105540	0.7798274	0.5608570	0.3350957	0.4366159	0.1906335
GBM_3_AutoML_20181212_105540	0.7786685	0.5617903	0.3255378	0.4371886	0.1911339
XGBoost_grid_1_AutoML_20181212_105540_model_2	0.7744105	0.5750165	0.3228112	0.4427003	0.1959836
GBM_4_AutoML_20181212_105540	0.7714260	0.5697120	0.3374203	0.4410703	0.1945430
GBM_grid_1_AutoML_20181212_105540_model_1	0.7697524	0.5725826	0.3443314	0.4424524	0.1957641
GBM_grid_1_AutoML_20181212_105540_model_2	0.7543664	0.9185673	0.3558550	0.4966377	0.2466490
DRF_1_AutoML_20181212_105540	0.7428924	0.5958832	0.3554027	0.4527742	0.2050045
XRT_1_AutoML_20181212_105540	0.7420910	0.5993457	0.3565826	0.4531168	0.2053148
DeepLearning_grid_1_AutoML_20181212_105540_model_2	0.7417952	0.6014974	0.3682910	0.4549035	0.2069372
XGBoost_grid_1_AutoML_20181212_105540_model_1	0.6935538	0.6207021	0.4058805	0.4657911	0.2169614
DeepLearning_1_AutoML_20181212_105540	0.6913704	0.6379538	0.4093513	0.4717801	0.2225765
DeepLearning_grid_1_AutoML_20181212_105540_model_1	0.6900835	0.6617941	0.4184695	0.4766352	0.2271811
GLM_grid_1_AutoML_20181212_105540_model_1	0.6826481	0.6385205	0.3972341	0.4726827	0.2234290

Example Leaderboard for binary classification (Higgs 10k)

H2O AutoML on Kaggle

Erin LeDell
@ledell

I had fun playing with @h2oai #AutoML on the #KaggleDaysSF hackathon today. One line of code, 8th place!

Ran H2O AutoML for 100 mins: it trained & 5-fold CV 43 models & 2 stacked ensembles. Wish I had joined the comp earlier & run longer! ⏰
Code here: gist.github.com/ledell/4d4cd24...

#	△pub	Team Name	Score	Entries	Last
1	▲30	Erkut & Mark	0.61691	12	2h
2	▲1	Google AutoML	0.61598	8	3h
3	▼2	Sweet Deal	0.61576	20	2h
4	▲11	Arno Candel @ H2O.ai	0.61549	17	2h
5	▼1	ALDAPOP	0.61504	11	2h
6	▲12	9hr Overfitness	0.61437	17	2h
7	▼5	Shlandryn	0.61413	38	2h
8	▲2	Erin (H2O AutoML 100 mins)	0.61312	5	2h
9	▼2	[ods.ai] bestfitting	0.61298	27	2h
10	▲18	We are not Pavel Pleskov	0.61237	30	2h
11	▲12	Super Organic	0.61222	16	3h
12	▲7	ryches	0.61210	7	3h
13	▼5	City.AI	0.61209	26	2h
14	▼5	ALGCAMCHI	0.61200	25	2h
15	▲6	underfit	0.61047	28	2h
16	▲17	International Triangle	0.60984	17	2h
17	▲25	PaulHassamRainier	0.60980	17	2h
18	▼2	Dmitry Larko	0.60968	10	2h
19	▼8	Keep Getting Worse	0.60959	25	2h
20	▲20	Brendan Borin	0.60959	9	2h
21	▲3	Michael Maguire	0.60958	3	2h
22	▲12	RoseKnight401	0.60957	22	2h
23	▲14	Pavel Pleskov	0.60942	5	2h
24	▲2	it's not working anymore	0.60900	28	2h
25	▼20	JVS	0.60864	20	2h
26	▼9	mrinzy	0.60860	12	2h

7:48 PM · Apr 11, 2019 · Twitter Web Client

View Tweet activity

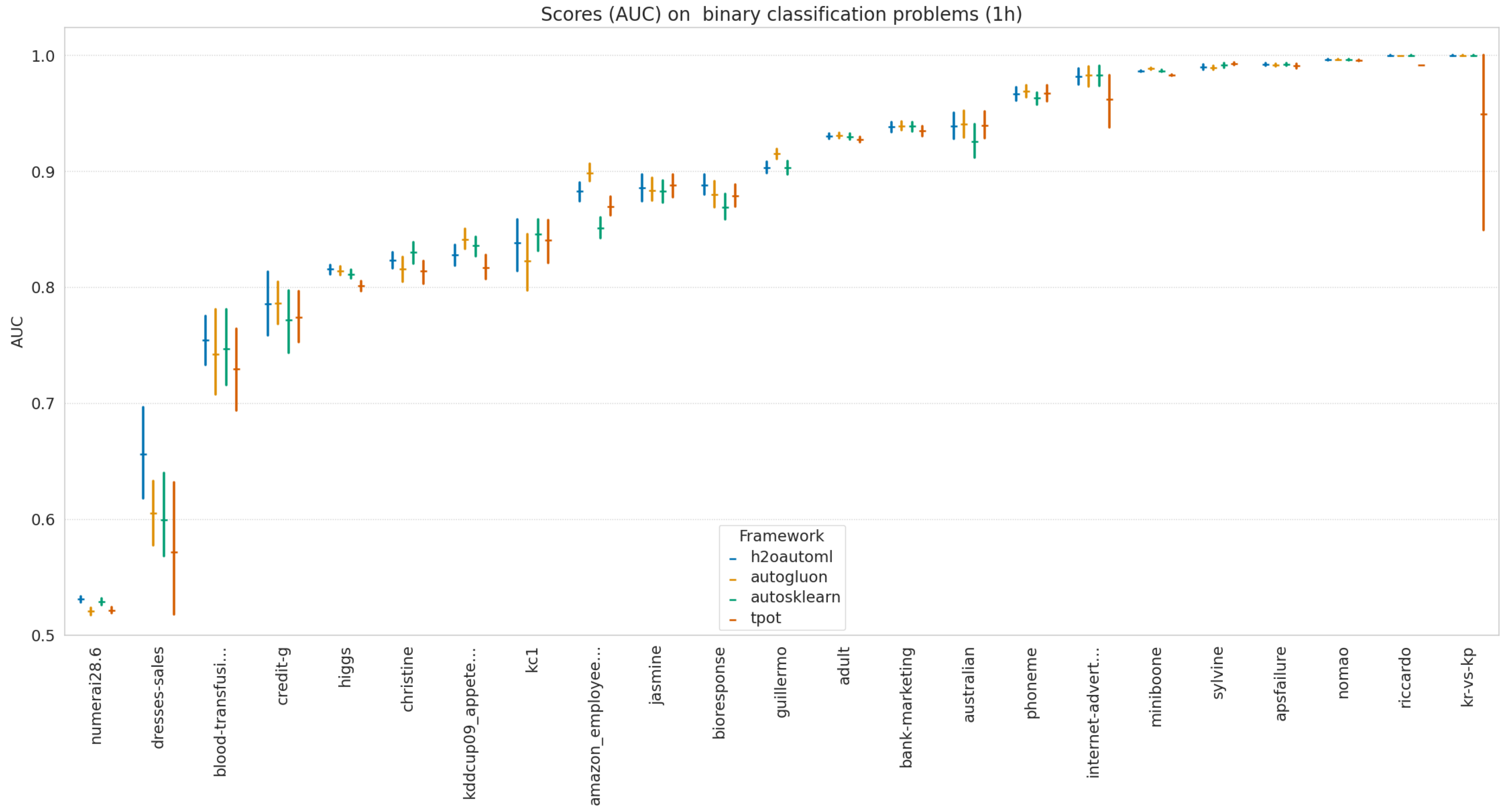
36 Retweets **198** Likes

- On a one-day Kaggle competition, AutoML tools ruled the leaderboard.
- In regular competitions, teams have much more time to do sophisticated, hand-crafted feature engineering, and so the humans are still winning (for now... 😈).

https://twitter.com/ledell/status/1116533416155963392

H2O AutoML paper

The H2O AutoML paper was accepted at ICML 2020 AutoML Workshop



- *Official H2O AutoML paper*
- *Algorithm details*
- *Scalability study (10k - 100M rows)*
- *OpenML AutoML Benchmark results*

ICML paper ⬇️ https://tinyurl.com/icml2020-h2o-paper

Learn H2O AutoML!

- User Guide: https://tinyurl.com/h2o-automl-docs
- R & Py tutorials: https://tinyurl.com/h2o-automl-tutorials
- LatinR tutorial: https://tinyurl.com/latinr-h2o
- useR! 2020: https://github.com/ledell/useR2020-automl

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
- Events & Meetups: http://h2o.ai/events

Thank you!

@ledell on Github, Twitter
erin@h2o.ai