
H2O.ai
AutoML Vs Fully Automatic: Comparisons between H2O-3 and Driverless AI
Tuesday, December 10, 6 PM - 8 PM
New York
Thomas Ott - Solutions Engineer
Meetup

Gratitude

H2O.ai Overview

Company	Founded in Silicon Valley in 2012 Funded: $147.5M Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures, CapitalOne and most recently $72.5M led by Goldman Sachs
Products	• H2O Open Source Machine Learning (18,000 organizations) • H2O Driverless AI – Automatic Machine Learning
Team	200+ AI experts (Expert data scientists, Kaggle Grandmasters, Distributed Computing, Visualization)
Global	Mountain View, NYC, Toronto, London, Prague, India

Global Customers

Global Industrial/ Agriculture | Telcos | Media and Marketing | Retail | Financial | Insurance | Advisory, Accounting and Government | Healthcare

H2O.ai Product Suite

Open Source

H2O	Spark + H2O SPARKLING WATER	H2O4GPU
In-memory, distributed machine learning algorithms with H2O Flow GUI	H2O AI open source engine integration with Spark	Lightning fast machine learning on GPUs

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise support subscriptions

DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

- Enterprise software
- Built for domain users, analysts and data scientists – GUI-based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

H2O.ai Product Suite

Open Source

H2O

In-memory, distributed machine learning algorithms with H2O Flow GUI

Spark + H2O SPARKLING WATER

H2O AI open source engine integration with Spark

H2O4GPU

Lightning fast machine learning on GPUs

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise support subscriptions

DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

- Enterprise software
- Built for domain users, analysts and data scientists – GUI-based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- User licenses on a per seat basis (annual subscription)

H2O-3's AutoML

What's in the AutoML box?

- Works with Python / R / Java / H2O Flow
- Horizontally Scalable* and Time Based
- Hyperparameter Tuning (Random Grid)
- Algorithms:
  - XGBoostGBM
  - GBM
  - Deep Learning
  - Distributed Random Forests
  - GLM
  - Stacked Ensemble

** XGBoostGBM is turned off by default for multi-node*

What's in the AutoML box?

- Works with Python / R / Java / **H2O Flow << Demo Tonight**
- Horizontally Scalable* and Time Based
- Hyperparameter Tuning (Random Grid)
- Algorithms
  - XGBoostGBM
  - GBM
  - Deep Learning
  - Distributed Random Forests
  - GLM
  - Stacked Ensemble

** XGBoostGBM is turned off by default for multi-node*

What's in the AutoML box?

- Works with Python / R / Java / **H2O Flow << Demo Tonight**
- Horizontally Scalable* and Time Based
- Hyperparameter Tuning (Random Grid)
- Algorithms
  - XGBoostGBM
  - GBM
  - Deep Learning
  - Distributed Random Forests
  - GLM
  - Stacked Ensemble

MOJO

** XGBoostGBM is turned off by default for multi-node*

What's in the AutoML box?

H2O's AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.

What's in the AutoML box?

H2O's AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.

Driverless AI

What's in the Driverless AI box?

- GUI based / Python & R Clients
- Vertically Scaling / CPU & GPU based
- Hyperparameter Tuning (Evolutionary as part of GA)
- Feature Engineering / AutoDoc / Machine Learning Interpretability (Shapley / LIME)
- Algorithms:
  - XGBoostGBM / GBM
  - LightGBM
  - GLM
  - TensorFlow (CNN)
  - FTRL
  - Stacked Ensemble

What's in the Driverless AI box?

- GUI based / Python & R Clients / BYOR
- Vertically Scaling / CPU & GPU based
- Hyperparameter Tuning (Evolutionary as part of GA)
- Feature Engineering / AutoDoc / Machine Learning Interpretability (Shapley / LIME)
- Algorithms:
  - XGBoostGBM / GBM
  - LightGBM
  - GLM
  - TensorFlow (CNN)
  - FTRL
  - Stacked Ensemble

Python
MOJO Java
MOJO C++
1 Click Deploy

Kaggle Dataset

Loan Loss Default

Variable	Description
Id	A unique identifier associated with an application.
F1	Numerical features
…	
F778	Numerical features
Loss	Target (numerical) / Optimize for M.A.E.

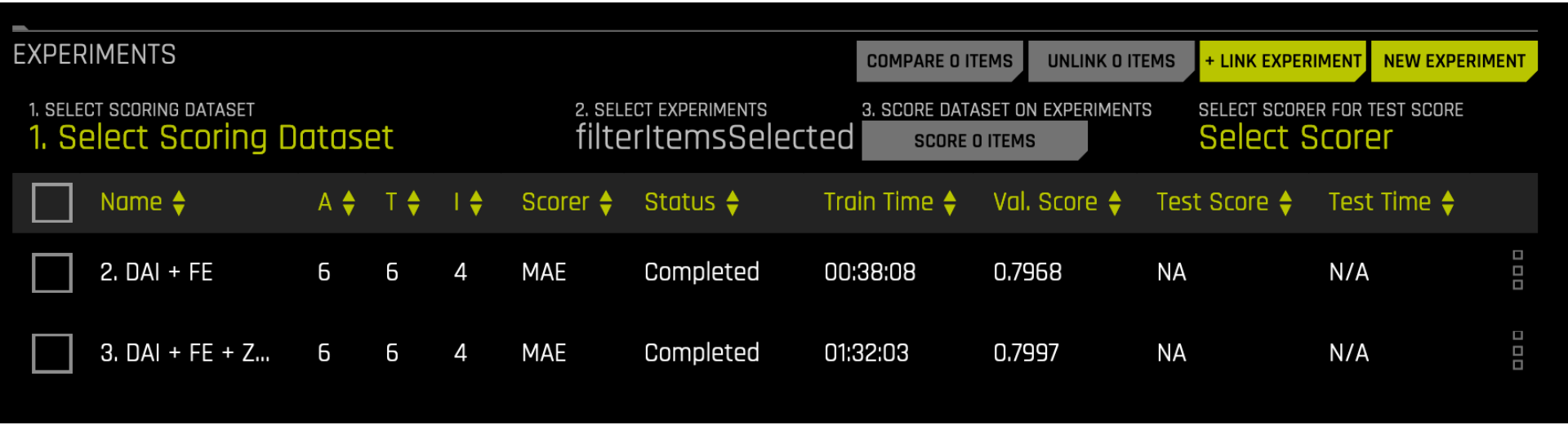
Reference: https://www.kaggle.com/c/loan-default-prediction

Driverless Experiments

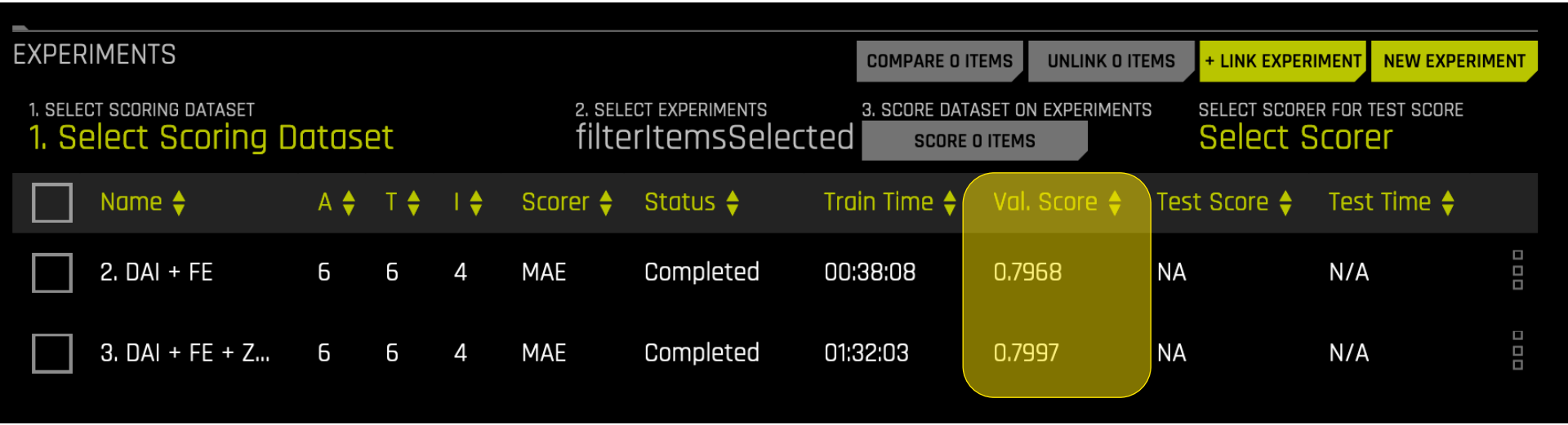
Driverless Experiments

- Generate 2 Experiments
  - Experiment #1: 6/6/4 Settings
  - Experiment #2: 6/6/4 Settings + BYOR
- Run on 1 GPU / Approximately 30 minutes to 1.5 hours
- Compare Features Side by Side

Driverless Experiments



Driverless Experiments



Driverless Experiments - Variable Importance

H2O.ai

VARIABLE IMPORTANCE	
783_ClusterDist50:f170:f517:f766.5	1.00
789_NumToCatTE:f471:f630.0	0.61
783_ClusterDist50:f170:f517:f766.4	0.56
783_ClusterDist50:f170:f517:f766.7	0.53
783_ClusterDist50:f170:f517:f766.6	0.52
783_ClusterDist50:f170:f517:f766.12	0.43
783_ClusterDist50:f170:f517:f766.10	0.27
380_f46	0.24
783_ClusterDist50:f170:f517:f766.14	0.22
407_f493	0.16
783_ClusterDist50:f170:f517:f766.9	0.13
783_ClusterDist50:f170:f517:f766.15	0.06
306_f39	0.03

VARIABLE IMPORTANCE	
689_f766	1.00
388_f471	0.93
594_f674	0.88
321_f404	0.85
590_f670	0.84
589_f67	0.82
727_CVTE:f137.0	0.75
731_CVTE:f276.0	0.72
107_f2	0.71
453_f536	0.68
732_CVTE:f277.0	0.67
547_f630	0.65
238_f322	0.64
32_f13	0.63
0_f1	0.62

H2O-3 AutoML Experiment

H2O-3 AutoML – 4 Hour Training

▾ MODELS

models sorted in order of mean_residual_deviance, best first

	model_id	mean_residual_deviance	rmse	mse	mae
0	StackedEnsemble_AllModels_AutoML_20191205_134903	18.457724565363122	4.296245403298457	18.457724565363122	1.4360225136135123
1	StackedEnsemble_BestOfFamily_AutoML_20191205_134903	18.488463419168774	4.2998213240981045	18.488463419168774	1.4398850252912432
2	XGBoost_grid_1_AutoML_20191205_134903_model_6	18.524848729415798	4.304050270316995	18.524848729415798	1.4084007958468263
3	XGBoost_grid_1_AutoML_20191205_134903_model_20	18.531881298735822	4.304867163889709	18.531881298735822	1.3932624527912796
4	XGBoost_grid_1_AutoML_20191205_134903_model_11	18.532076903555605	4.304889882860606	18.532076903555605	1.3443163254323498
5	XGBoost_grid_1_AutoML_20191205_134903_model_21	18.533253687333488	4.305026560583975	18.533253687333488	1.414472436842969
6	GBM_grid_1_AutoML_20191205_134903_model_3	18.55352115412073	4.3073798479029834	18.55352115412073	1.4391232532568006
7	XGBoost_grid_1_AutoML_20191205_134903_model_2	18.577092294616232	4.310115113847452	18.577092294616232	1.4232732504339223

H2O-3 AutoML – 4 Hour Training

▾ MODELS

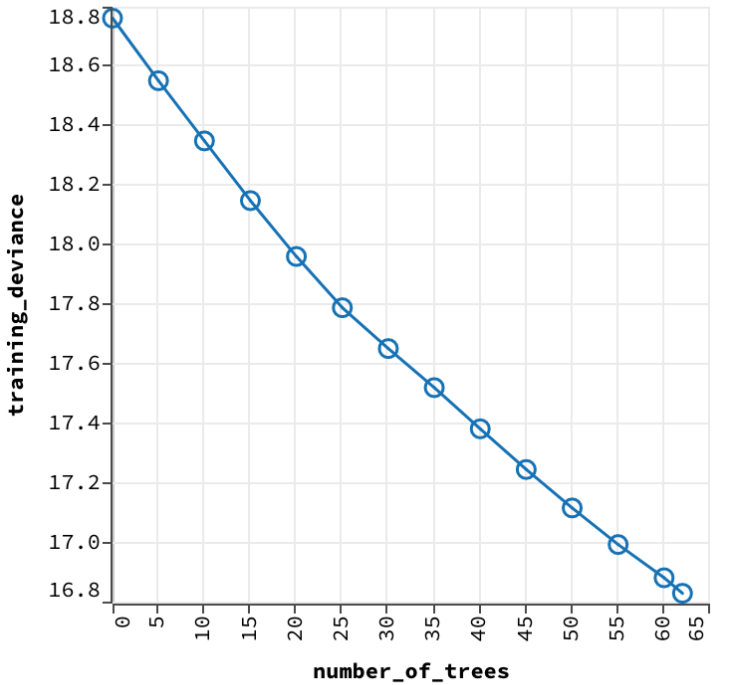
models sorted in order of mean_residual_deviance, best first

	model_id	mean_residual_deviance	rmse	mse	mae
0	StackedEnsemble_AllModels_AutoML_20191205_134903	18.457724565363122	4.296245403298457	18.457724565363122	1.4360225136135123
1	StackedEnsemble_BestOfFamily_AutoML_20191205_134903	18.488463419168774	4.2998213240981045	18.488463419168774	1.4398850252912432
2	XGBoost_grid_1_AutoML_20191205_134903_model_6	18.524848729415798	4.304050270316995	18.524848729415798	1.4084007958468263
3	XGBoost_grid_1_AutoML_20191205_134903_model_20	18.531881298735822	4.304867163889709	18.531881298735822	1.3932624527912796
4	XGBoost_grid_1_AutoML_20191205_134903_model_11	18.532076903555605	4.304889882860606	18.532076903555605	1.3443163254323498
5	XGBoost_grid_1_AutoML_20191205_134903_model_21	18.533253687333488	4.305026560583975	18.533253687333488	1.414472436842969
6	GBM_grid_1_AutoML_20191205_134903_model_3	18.55352115412073	4.3073798479029834	18.55352115412073	1.4391232532568006
7	XGBoost_grid_1_AutoML_20191205_134903_model_2	18.577092294616232	4.310115113847452	18.577092294616232	1.4232732504339223

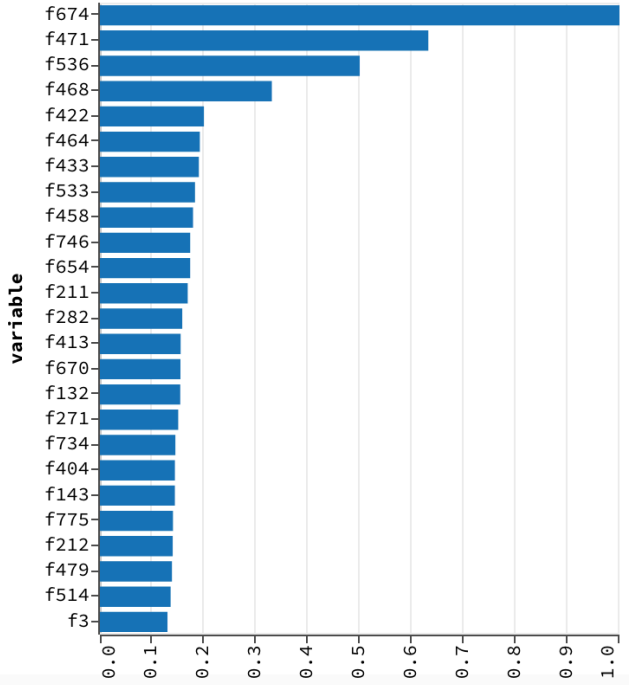
MAE = 1.34 using XGBoost model #11

H2O-3 AutoML – Features

SCORING HISTORY - DEVIANCE

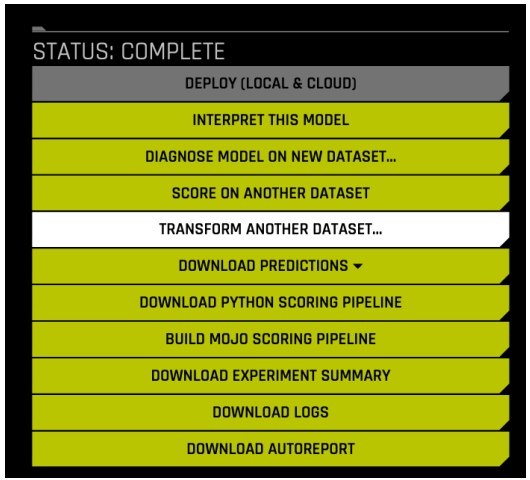


VARIABLE IMPORTANCES



Using Driverless AI Features in AutoML

Transform Dataset into Transformed Features



	A	B	C	D	E	F	G	H	I
1	306_f39	380_f46	407_f493	783_ClusterDist50:f170:f517:f766.5	783_ClusterDist50:f170:f517:f766.6	783_ClusterDist50:f170:f517:f766.7	783_ClusterDist50:f170:f517:f766.9	783_ClusterDist50:f170:f517:f766.10	loss
2	0.65705	0.70127	19.72	4.1890683	3.468733	4.046106	3.430209	0	0
3	0.7471	0.80819	124	3.0490234	3.707606	2.2169568	3.531153	2.4769564	0
4	0.77405	0.8207	903	5.2391896	3.156621	1.0810538	5.9418473	4.172575	0
5	0.78385	0.86382	130.94	4.650567	2.7428029	1.2893546	5.1386466	3.2553873	0
6	0.79085	0.82485	399	1.9060979	4.804308	3.0140483	2.7850735	2.9816196	0
7	0.7269	0.89431	836.75	3.711679	3.1023474	2.0080702	4.0702352	2.6610246	1
8	0.76075	0.92268	252.92	2.8794038	4.3365793	2.0285404	3.8632266	3.372635	0
9	0.7995	0.88271	82	4.058473	3.7201104	1.4652498	4.8899136	3.8527265	0
10	0.7868	0.84464	82.5	4.264527	2.3753636	2.4006875	4.2668266	2.2144604	0

Load into AutoML and Run for 1 Hour

H2O.ai

▾ MODELS

models sorted in order of mean_residual_deviance, best first

32	GBM_4_AutoML_20191205_170249	19.53685310518864	4.420051255945867	19.53685310518864	1.452662850257629	NaN
33	GBM_1_AutoML_20191205_170249	19.53887942185216	4.420280468686593	19.53887942185216	1.4491605755184283	NaN
34	XGBoost_grid_1_AutoML_20191205_170249_model_1	19.550453470904767	4.421589473357377	19.550453470904767	1.435419096811673	NaN
35	GBM_grid_1_AutoML_20191205_170249_model_11	19.560374134017135	4.422711174609657	19.560374134017135	1.4618233503803184	0.73
36	XGBoost_grid_1_AutoML_20191205_170249_model_19	19.574376138075163	4.424293857563618	19.574376138075163	1.4323884281055193	0.73
37	XGBoost_grid_1_AutoML_20191205_170249_model_17	19.574499892978885	4.424307843378316	19.574499892978885	1.4434372742349393	0.74
38	XGBoost_grid_1_AutoML_20191205_170249_model_12	19.582265382649492	4.425185350089812	19.582265382649492	1.443978908677506	NaN
39	XGBoost_grid_1_AutoML_20191205_170249_model_15	19.63858313079829	4.431544102318997	19.63858313079829	1.447825694990862	NaN
40	DeepLearning_grid_1_AutoML_20191205_170249_model_6	19.669011907915188	4.434975976024582	19.669011907915188	1.047967422168763	0.63
41	DeepLearning_grid_1_AutoML_20191205_170249_model_5	19.701332086128847	4.4386182631680375	19.701332086128847	1.2262382861044774	NaN
42	GBM_grid_1_AutoML_20191205_170249_model_10	19.72759121230986	4.441575307513074	19.72759121230986	1.4504047077809836	NaN

AutoML Results

MODELS

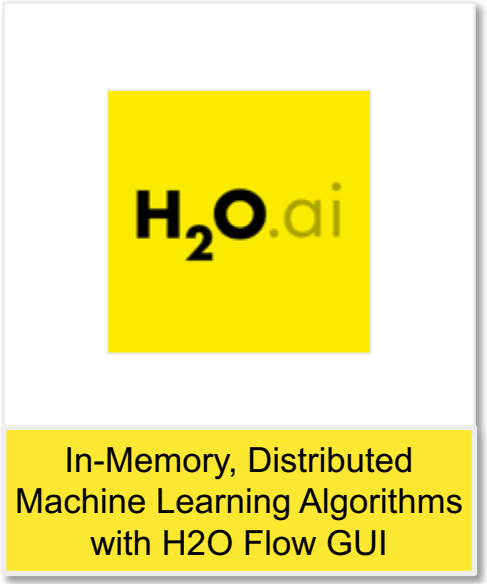
models sorted in order of mean_residual_deviance, best first

32	GBM_4_AutoML_20191205_170249	19.53685310518864	4.420051255945867	19.53685310518864	1.452662850257629	NaN
33	GBM_1_AutoML_20191205_170249	19.53887942185216	4.420280468686593	19.53887942185216	1.4491605755184283	NaN
34	XGBoost_grid_1_AutoML_20191205_170249_model_1	19.550453470904767	4.421589473357377	19.550453470904767	1.435419096811673	NaN
35	GBM_grid_1_AutoML_20191205_170249_model_11	19.560374134017135	4.422711174609657	19.560374134017135	1.4618233503803184	0.73
36	XGBoost_grid_1_AutoML_20191205_170249_model_19	19.574376138075163	4.424293857563618	19.574376138075163	1.4323884281055193	0.73
37	XGBoost_grid_1_AutoML_20191205_170249_model_17	19.574499892978885	4.424307843378316	19.574499892978885	1.4434372742349393	0.74
38	XGBoost_grid_1_AutoML_20191205_170249_model_12	19.582265382649492	4.425185350089812	19.582265382649492	1.443978908677506	NaN
39	XGBoost_grid_1_AutoML_20191205_170249_model_15	19.63858313079829	4.431544102318997	19.63858313079829	1.447825694990862	NaN
40	DeepLearning_grid_1_AutoML_20191205_170249_model_6	19.669011907915188	4.434975976024582	19.669011907915188	1.047967422168763	0.63
41	DeepLearning_grid_1_AutoML_20191205_170249_model_5	19.701332086128847	4.4386182631680375	19.701332086128847	1.2262382861044774	NaN
42	GBM_grid_1_AutoML_20191205_170249_model_10	19.72759121230986	4.441575307513074	19.72759121230986	1.4504047077809836	NaN

MAE = 1.04 using DL model_6

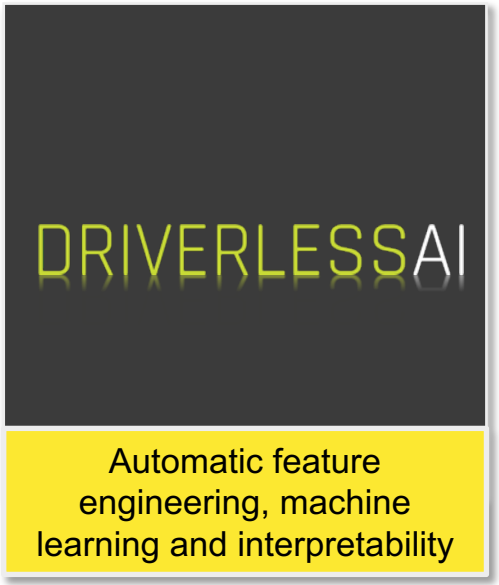
Reasons to Use

Reasons to use H2O-3 AutoML



- Open Source
- High Degree of User Control
- In Memory & Distributed
- Web UI, Python Client, R Client
- Autogenerated Java Scoring Pipeline

Reasons to use Driverless AI



- Commercial license
- High degree of automation
- Optimized for GPU's
- Web UI, Python Client, R Client
- Feature Engineering + BYOR
- Autogenerated Java/Python/C++ Scoring Pipelines
- MLI + MLI Scoring Pipeline
- AutoDoc

Thank You

Thomas Ott
Github: https://github.com/tomott12345/