Driving the Driverless AI

An Exercise based on Kaggle Competition

Jo-fai Chow - joe@h2o.ai

10-Feb-2018

Agenda

- Kaggle Competition and Dataset
- Baseline Model
- Kaggle Submission
- Using Driverless AI
    - AWS
    - Data Visualisation
    - Making Predictions
- **Goal**: Beating the Baseline without Manual Feature Engieering
- Beyond this Exercise

Kaggle Competition & Dataset

House Prices: Advanced Regression Techniques

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

3,877 teams · 2 years to go

Overview

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

Practice Skills

- Creative feature engineering
- Advanced regression techniques like random forest and gradient boosting

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

3,877 teams · 2 years to go

Overview

Learning with Kaggle Kernels

Kaggle Kernels is an in-browser computational environment that is fully integrated with most competition datasets. Kernels is preloaded with most data science packages and libraries. It supports scripts and Jupyter Notebooks in R and Python, as well as RMarkdown reports. You can create submission files with Kernels and also use it to explore the competition data.

To get started with Kernels you can either:

1. Create a new script or notebook on the Kernels tab or
2. "Fork" any kernel to create an editable copy for you to experiment with

We've selected some of the best kernels to help you get started with the competition. You can use the below kernels to create a submission file or to explore the data. Simply open the script or notebook and click "fork" to create an editable copy.

Getting started with R

Detailed Exploratory Data Analysis Using R

- Use RMarkdown and popular R packages like data.table, dplyr, and ggplot2
- Take an in-depth look at missing values, distributions, and correlations
- Click on the "Code" tab to see the underlying code which combines markdown text and R scripts

Fun with Real Estate Data

- Use Rmarkdown to learn advanced regression techniques like random forests and XGBoost

XGBoost with Parameter Tuning

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs ...

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

3,877 teams · 2 years to go

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Competition Data

- sample_submission.cs...
- test.csv
- train.csv
- data_description.txt
- sample_submission.cs...
- test.csv.gz
- train.csv.gz

data_description.txt 13.06 KB **Download**

Data Description

File descriptions

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data fields

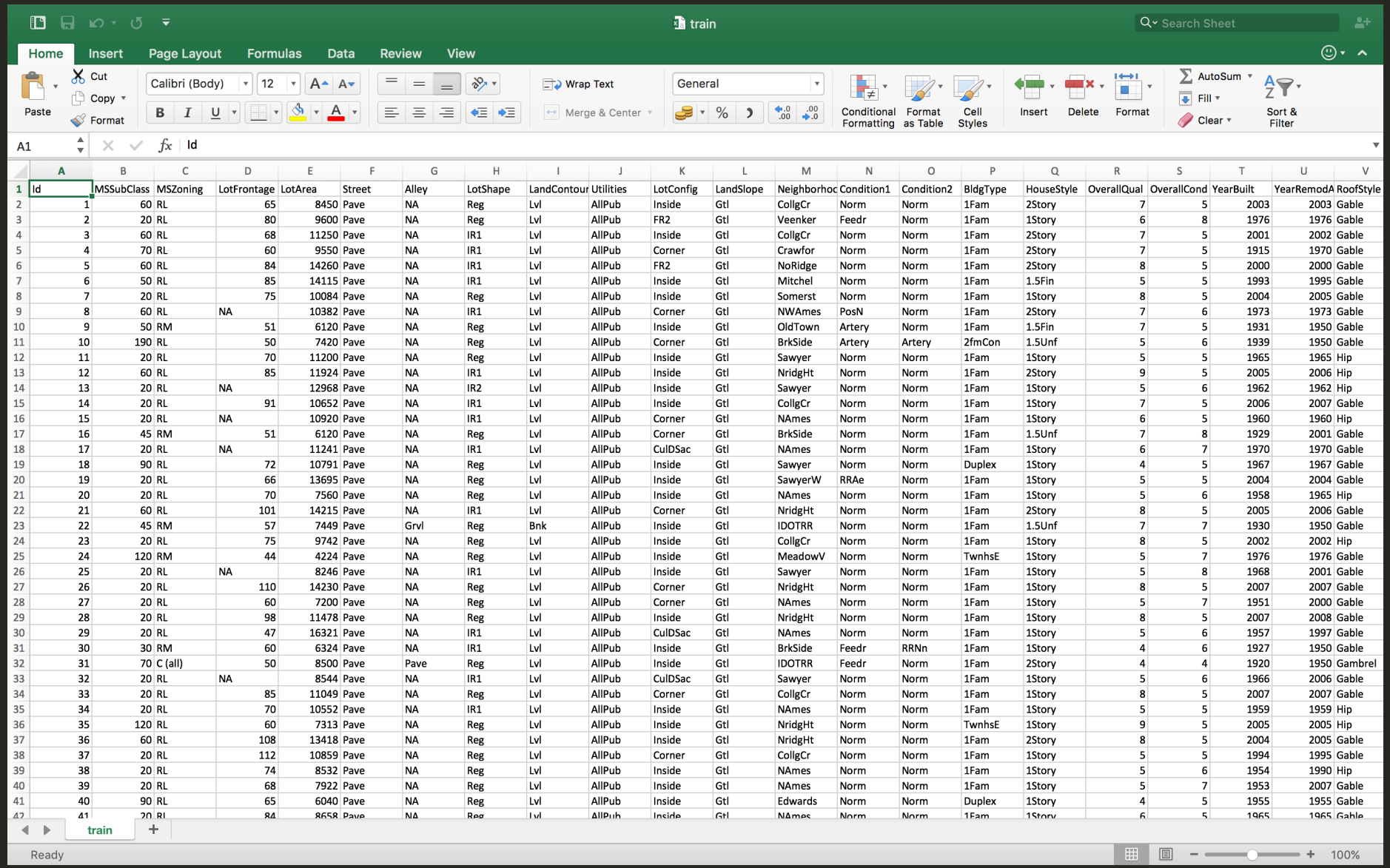
Here's a brief version of what you'll find in the data description file.

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass:** The building class
- **MSZoning:** The general zoning classification
- **LotFrontage:** Linear feet of street connected to property
- **LotArea:** Lot size in square feet
- **Street:** Type of road access
- **Alley:** Type of alley access
- **LotShape:** General shape of property
- **LandContour:** Flatness of the property
- **Utilities:** Type of utilities available
- **LotConfig:** Lot configuration
- **LandSlope:** Slope of property
- **Neighborhood:** Physical locations within Ames city limits
- **Condition1:** Proximity to main road or railroad
- **Condition2:** Proximity to main road or railroad (if a second is present)
- **BldgType:** Type of dwelling
- **HouseStyle:** Style of dwelling
- **OverallQual:** Overall material and finish quality
- **OverallCond:** Overall condition rating
- **YearBuilt:** Original construction date
- **YearRemodAdd:** Remodel date
- **RoofStyle:** Type of roof
- **RoofMatl:** Roof material
- **Exterior1st:** Exterior covering on house
- **Exterior2nd:** Exterior covering on house (if more than one material)
- **MasVnrType:** Masonry veneer type
- **MasVnrArea:** Masonry veneer area in square feet
- **ExterQual:** Exterior material quality
- **ExterCond:** Present condition of the material on the exterior

- **FullBath:** Full bathrooms above grade
- **HalfBath:** Half baths above grade
- **Bedroom:** Number of bedrooms above basement level
- **Kitchen:** Number of kitchens
- **KitchenQual:** Kitchen quality
- **TotRmsAbvGrd:** Total rooms above grade (does not include bathrooms)
- **Functional:** Home functionality rating
- **Fireplaces:** Number of fireplaces
- **FireplaceQu:** Fireplace quality
- **GarageType:** Garage location
- **GarageYrBlt:** Year garage was built
- **GarageFinish:** Interior finish of the garage
- **GarageCars:** Size of garage in car capacity
- **GarageArea:** Size of garage in square feet
- **GarageQual:** Garage quality
- **GarageCond:** Garage condition
- **PavedDrive:** Paved driveway
- **WoodDeckSF:** Wood deck area in square feet
- **OpenPorchSF:** Open porch area in square feet
- **EnclosedPorch:** Enclosed porch area in square feet
- **3SsnPorch:** Three season porch area in square feet
- **ScreenPorch:** Screen porch area in square feet
- **PoolArea:** Pool area in square feet
- **PoolQC:** Pool quality
- **Fence:** Fence quality
- **MiscFeature:** Miscellaneous feature not covered in other categories
- **MiscVal:** $Value of miscellaneous feature
- **MoSold:** Month Sold
- **YrSold:** Year Sold
- **SaleType:** Type of sale
- **SaleCondition:** Condition of sale

Dataset

- Source: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- ./data/
  - train.csv
  - test.csv
  - sample_submission.csv



train

Search Sheet

Home Insert Page Layout Formulas Data Review View

Paste Cut Copy Format

Calibri (Body) 12

Wrap Text Merge & Center

General

Conditional Formatting Format as Table Cell Styles

Insert Delete Format

AutoSum Fill Clear Sort & Filter

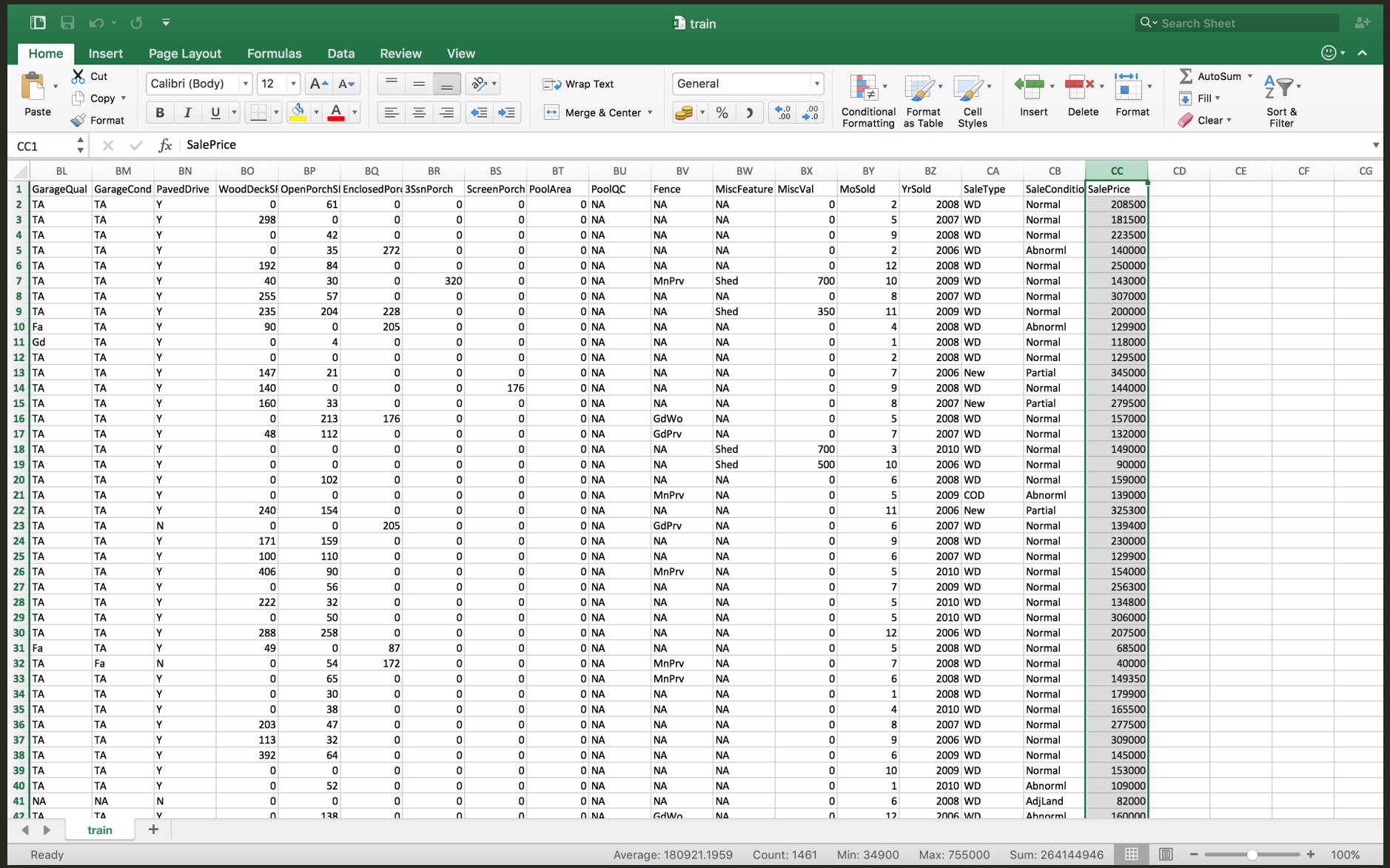
A1 Id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhoc	Condition1	Condition2	BldgType	HouseStyle	OverallQual	OverallCond	YearBuilt	YearRemodA	RoofStyle
2	1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2003	2003	Gable
3	2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	Veenker	Feedr	Norm	1Fam	1Story	6	8	1976	1976	Gable
4	3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	2Story	7	5	2001	2002	Gable
5	4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Crawfor	Norm	Norm	1Fam	2Story	7	5	1915	1970	Gable
6	5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2	Gtl	NoRidge	Norm	Norm	1Fam	2Story	8	5	2000	2000	Gable
7	6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Mitchel	Norm	Norm	1Fam	1.5Fin	5	5	1993	1995	Gable
8	7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	Norm	1Fam	1Story	8	5	2004	2005	Gable
9	8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NWAmes	PosN	Norm	1Fam	2Story	7	6	1973	1973	Gable
10	9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	OldTown	Artery	Norm	1Fam	1.5Fin	7	5	1931	1950	Gable
11	10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery	Artery	2fmCon	1.5Unf	5	6	1939	1950	Gable
12	11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	5	1965	1965	Hip
13	12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	9	5	2005	2006	Hip
14	13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1962	1962	Hip
15	14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	7	5	2006	2007	Gable
16	15	20	RL	NA	10920	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	1960	1960	Hip
17	16	45	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5Unf	7	8	1929	2001	Gable
18	17	20	RL	NA	11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	7	1970	1970	Gable
19	18	90	RL	72	10791	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	Duplex	1Story	4	5	1967	1967	Gable
20	19	20	RL	66	13695	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	SawyerW	RRAe	Norm	1Fam	1Story	5	5	2004	2004	Gable
21	20	20	RL	70	7560	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1958	1965	Hip
22	21	60	RL	101	14215	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	2005	2006	Gable
23	22	45	RM	57	7449	Pave	Grvl	Reg	Bnk	AllPub	Inside	Gtl	IDOTRR	Norm	Norm	1Fam	1.5Unf	7	7	1930	1950	Gable
24	23	20	RL	75	9742	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	2002	2002	Hip
25	24	120	RM	44	4224	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	MeadowV	Norm	Norm	TwnhsE	1Story	5	7	1976	1976	Gable
26	25	20	RL	NA	8246	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	8	1968	2001	Gable
27	26	20	RL	110	14230	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2007	2007	Gable
28	27	20	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	1951	2000	Gable
29	28	20	RL	98	11478	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	1Story	8	5	2007	2008	Gable
30	29	20	RL	47	16321	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1957	1997	Gable
31	30	30	RM	60	6324	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr	RRNn	1Fam	1Story	4	6	1927	1950	Gable
32	31	70	C (all)	50	8500	Pave	Pave	Reg	Lvl	AllPub	Inside	Gtl	IDOTRR	Feedr	Norm	1Fam	2Story	4	4	1920	1950	Gambrel
33	32	20	RL	NA	8544	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	Sawyer	Norm	Norm	1Fam	1Story	5	6	1966	2006	Gable
34	33	20	RL	85	11049	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	1Story	8	5	2007	2007	Gable
35	34	20	RL	70	10552	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	5	1959	1959	Hip
36	35	120	RL	60	7313	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	TwnhsE	1Story	9	5	2005	2005	Hip
37	36	60	RL	108	13418	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NridgHt	Norm	Norm	1Fam	2Story	8	5	2004	2005	Gable
38	37	20	RL	112	10859	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	CollgCr	Norm	Norm	1Fam	1Story	5	5	1994	1995	Gable
39	38	20	RL	74	8532	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	6	1954	1990	Hip
40	39	20	RL	68	7922	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	5	7	1953	2007	Gable
41	40	90	RL	65	6040	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	Duplex	1Story	4	5	1955	1955	Gable
42	41	20	RL	84	8658	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Story	6	5	1965	1965	Gable

train

Ready

100%



train

Search Sheet

Home Insert Page Layout Formulas Data Review View

Paste Cut Copy Format

Calibri (Body) 12

Wrap Text Merge & Center

General

Conditional Formatting Format as Table Cell Styles

Insert Delete Format

AutoSum Fill Clear

Sort & Filter

CC1 SalePrice

	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG
1	GarageQual	GarageCond	PavedDrive	WoodDeckSf	OpenPorchSf	EnclosedPor	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleConditio	SalePrice				
2	TA	TA	Y	0	61	0	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	208500				
3	TA	TA	Y	298	0	0	0	0	0	NA	NA	NA	0	5	2007	WD	Normal	181500				
4	TA	TA	Y	0	42	0	0	0	0	NA	NA	NA	0	9	2008	WD	Normal	223500				
5	TA	TA	Y	0	35	272	0	0	0	NA	NA	NA	0	2	2006	WD	Abnorml	140000				
6	TA	TA	Y	192	84	0	0	0	0	NA	NA	NA	0	12	2008	WD	Normal	250000				
7	TA	TA	Y	40	30	0	320	0	0	NA	MnPrv	Shed	700	10	2009	WD	Normal	143000				
8	TA	TA	Y	255	57	0	0	0	0	NA	NA	NA	0	8	2007	WD	Normal	307000				
9	TA	TA	Y	235	204	228	0	0	0	NA	NA	Shed	350	11	2009	WD	Normal	200000				
10	Fa	TA	Y	90	0	205	0	0	0	NA	NA	NA	0	4	2008	WD	Abnorml	129900				
11	Gd	TA	Y	0	4	0	0	0	0	NA	NA	NA	0	1	2008	WD	Normal	118000				
12	TA	TA	Y	0	0	0	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	129500				
13	TA	TA	Y	147	21	0	0	0	0	NA	NA	NA	0	7	2006	New	Partial	345000				
14	TA	TA	Y	140	0	0	0	176	0	NA	NA	NA	0	9	2008	WD	Normal	144000				
15	TA	TA	Y	160	33	0	0	0	0	NA	NA	NA	0	8	2007	New	Partial	279500				
16	TA	TA	Y	0	213	176	0	0	0	NA	GdWo	NA	0	5	2008	WD	Normal	157000				
17	TA	TA	Y	48	112	0	0	0	0	NA	GdPrv	NA	0	7	2007	WD	Normal	132000				
18	TA	TA	Y	0	0	0	0	0	0	NA	NA	Shed	700	3	2010	WD	Normal	149000				
19	TA	TA	Y	0	0	0	0	0	0	NA	NA	Shed	500	10	2006	WD	Normal	90000				
20	TA	TA	Y	0	102	0	0	0	0	NA	NA	NA	0	6	2008	WD	Normal	159000				
21	TA	TA	Y	0	0	0	0	0	0	NA	MnPrv	NA	0	5	2009	COD	Abnorml	139000				
22	TA	TA	Y	240	154	0	0	0	0	NA	NA	NA	0	11	2006	New	Partial	325300				
23	TA	TA	N	0	0	205	0	0	0	NA	GdPrv	NA	0	6	2007	WD	Normal	139400				
24	TA	TA	Y	171	159	0	0	0	0	NA	NA	NA	0	9	2008	WD	Normal	230000				
25	TA	TA	Y	100	110	0	0	0	0	NA	NA	NA	0	6	2007	WD	Normal	129900				
26	TA	TA	Y	406	90	0	0	0	0	NA	MnPrv	NA	0	5	2010	WD	Normal	154000				
27	TA	TA	Y	0	56	0	0	0	0	NA	NA	NA	0	7	2009	WD	Normal	256300				
28	TA	TA	Y	222	32	0	0	0	0	NA	NA	NA	0	5	2010	WD	Normal	134800				
29	TA	TA	Y	0	50	0	0	0	0	NA	NA	NA	0	5	2010	WD	Normal	306000				
30	TA	TA	Y	288	258	0	0	0	0	NA	NA	NA	0	12	2006	WD	Normal	207500				
31	Fa	TA	Y	49	0	87	0	0	0	NA	NA	NA	0	5	2008	WD	Normal	68500				
32	TA	Fa	N	0	54	172	0	0	0	NA	MnPrv	NA	0	7	2008	WD	Normal	40000				
33	TA	TA	Y	0	65	0	0	0	0	NA	MnPrv	NA	0	6	2008	WD	Normal	149350				
34	TA	TA	Y	0	30	0	0	0	0	NA	NA	NA	0	1	2008	WD	Normal	179900				
35	TA	TA	Y	0	38	0	0	0	0	NA	NA	NA	0	4	2010	WD	Normal	165500				
36	TA	TA	Y	203	47	0	0	0	0	NA	NA	NA	0	8	2007	WD	Normal	277500				
37	TA	TA	Y	113	32	0	0	0	0	NA	NA	NA	0	9	2006	WD	Normal	309000				
38	TA	TA	Y	392	64	0	0	0	0	NA	NA	NA	0	6	2009	WD	Normal	145000				
39	TA	TA	Y	0	0	0	0	0	0	NA	NA	NA	0	10	2009	WD	Normal	153000				
40	TA	TA	Y	0	52	0	0	0	0	NA	NA	NA	0	1	2010	WD	Abnorml	109000				
41	NA	NA	N	0	0	0	0	0	0	NA	NA	NA	0	6	2008	WD	AdjLand	82000				
42	TA	TA	Y	0	138	0	0	0	0	NA	GdWo	NA	0	12	2006	WD	Abnorml	160000				

train

Ready Average: 180921.1959 Count: 1461 Min: 34900 Max: 755000 Sum: 264144946 100%

Key to Success

Data Science Skills

- Creative Feature Engineering
- Advanced Regression Techniques
- Hyperparameters Tuning
- Models Stacking
- (months and years of practice)

Domain Knowledge

- Feature Engineering

Key to Success

Data Science Skills

- Creative Feature Engineering
- Advanced Regression Techniques
- Hyperparameters Tuning
- Models Stacking
- (months and years of practice)

Shortcuts

- Automatic Machine Learning
- H2O Driverless AI

Domain Knowledge

- Feature Engineering

Evaluation

Root Mean Squared Logarithmic Error (RMSLE)

Note: instead of RMSE as described on Evaluation page

Reference: https://www.slideshare.net/KhorSoonHin/rmsle-cost-function

Clip slide

Cost Functions

Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root Mean Squared Log Error (RMSLE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

prediction

actual

Clip slide

Alternative: RMSLE Intuition

RMSLE: log(Pi + 1) - log(Ai +1) = log((Pi + 1)/(Ai +1))

Only the percentual differences matter!

For example for P = 1000 and A = 500 would give you the roughly same error as when P = 100000 and A = 50000

RMSLE is usually used when you don't want to penalize huge differences in the predicted and true values when both predicted and true values are huge numbers.

Credits to Katerina Malahova for sharing this

Baseline Model

H2O Random Forest with Default Settings & 5-fold Cross-Validation

https://github.com/woobe/dai_oxford/blob/master/baseline.ipynb

In [1]:
```
# Import modules
import pandas as pd
import h2o
```

In [2]:
```
# Start a local H2O cluster
h2o.init(nthreads = -1)
```

```
Checking whether there is an H2O instance running at http://localhost:54321..... not found.
Attempting to start a local H2O server...
  Java Version: java version "1.8.0_72"; Java(TM) SE Runtime Environment (build 1.8.0_72-b15); Jav
a HotSpot(TM) 64-Bit Server VM (build 25.72-b15, mixed mode)
  Starting server from /Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/backend/bin/h2o.j
ar
  Ice root: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpho0r2p20
  JVM stdout: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpho0r2p20/h2o_jofaichow_started_f
rom_python.out
  JVM stderr: /var/folders/4z/p7yt7_4n4fj1jlyq6g4qhfbw0000gn/T/tmpho0r2p20/h2o_jofaichow_started_f
rom_python.err
  Server is running at http://127.0.0.1:54321
Connecting to H2O server at http://127.0.0.1:54321... successful.
Warning: Your H2O cluster version is too old (4 months)! Please download and install the latest ve
rsion from http://h2o.ai/download/
```

H2O cluster uptime:	01 secs
H2O cluster version:	3.14.0.6
H2O cluster version age:	4 months !!!
H2O cluster name:	H2O_from_python_jofaichow_6rld0a
H2O cluster total nodes:	1
H2O cluster free memory:	3.556 Gb
H2O cluster total cores:	8
H2O cluster allowed cores:	8
H2O cluster status:	accepting new members, healthy
H2O connection url:	http://127.0.0.1:54321
H2O connection proxy:	None
H2O internal security:	False
H2O API Extensions:	XGBoost, Algos, AutoML, Core V3, Core V4
Python version:	3.6.1 final

In [3]:
```
# Import data as H2O Data Frame
h_train = h2o.import_file("./data/train.csv")
h_test = h2o.import_file("./data/test.csv")
```

```
Parse progress: |████████████████████████████████████████████████████████| 100%
Parse progress: |████████████████████████████████████████████████████████| 100%
```

In [4]:
```
# Define features (or predictors) - notes: exclude 'Id' and target 'SalePrice'
features = h_train.col_names
features.remove('Id')
features.remove('SalePrice')
features
```

Out[4]:
```
['MSSubClass',
 'MSZoning',
 'LotFrontage',
 'LotArea',
 'Street',
 'Alley',
 'LotShape',
 'LandContour',
 'Utilities',
 'LotConfig',
 'LandSlope',
 'Neighborhood',
 'Condition1',
 'Condition2',
 'BldgType',
 'HouseStyle',
 'OverallQual',
 'OverallCond',
 'YearBuilt',
 'YearRemodAdd',
 'RoofStyle',
 'RoofMatl',
 'Exterior1st',
 'Exterior2nd',
 'MasVnrType',
 'MasVnrArea',
 'ExterQual',
 'ExterCond',
 'Foundation',
 'BsmtQual',
 'BsmtCond',
 'BsmtExposure',
 'BsmtFinType1',
 'BsmtFinSF1',
 'BsmtFinType2',
 'BsmtFinSF2',
 'BsmtUnfSF',
 'TotalBsmtSF',
 'Heating',
 'HeatingQC'
```

```
# Add a seed for reproducibility
drf_baseline = H2ORandomForestEstimator(model_id = 'drf_default',
                                        nfolds = 5,
                                        seed = 1234)

# Use .train() to build the model
drf_baseline.train(x = features,
                   y = 'SalePrice',
                   training_frame = h_train)
```

```
drf Model Build progress: |████████████████████████████████████████████████| 100%
```

In [6]:
```
# Model Summary
drf_baseline
```

```
Model Details
=============
H2ORandomForestEstimator :  Distributed Random Forest
Model Key:  drf_default


ModelMetricsRegression: drf
** Reported on train data. **

MSE: 837934511.4267565
RMSE: 28947.098497548188
MAE: 16824.085178911897
RMSLE: 0.13965617775881853
Mean Residual Deviance: 837934511.4267565

ModelMetricsRegression: drf
** Reported on cross-validation data. **

MSE: 814775750.9682842
RMSE: 28544.277026547446
MAE: 17009.127417487158
RMSLE: 0.14037094407475256
Mean Residual Deviance: 814775750.9682842
Cross-Validation Metrics Summary:
```

	mean	sd	cv_1_valid	cv_2_valid	cv_3_valid
mae	17035.605	1075.0117	15295.1875000	15951.877	16214.911
mean_residual_deviance	816360580.0000000	150756736.0000000	632738750.0000000	747103230.0000000	616277820.0000(
mse	816360580.0000000	150756736.0000000	632738750.0000000	747103230.0000000	616277820.0000(
r2	0.8717843	0.0165829	0.8779815	0.8840764	0.8959165
residual_deviance	816360580.0000000	150756736.0000000	632738750.0000000	747103230.0000000	616277820.0000(
rmse	28345.225	2540.5469	25154.299	27333.19	24824.943
rmsle	0.1396314	0.0126747	0.1264608	0.1210312	0.1289151

In [7]:
```
# Make predictions
yhat_test = drf_baseline.predict(h_test)
yhat_test = yhat_test.as_data_frame()
```

```
drf prediction progress: |████████████████████████████████████████████████| 100%
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'MSZoning' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'Utilities' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'Exterior1st' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'Exterior2nd' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'KitchenQual' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'Functional' has levels not trained on: [NA]
  warnings.warn(w)
/Users/jofaichow/anaconda/lib/python3.6/site-packages/h2o/job.py:69: UserWarning: Test/Validation
dataset column 'SaleType' has levels not trained on: [NA]
  warnings.warn(w)
```

In [8]:
```
# Create Kaggle Submission
d_sub = pd.read_csv("./data/sample_submission.csv")
d_sub = pd.concat([d_sub['Id'].reset_index(drop=True), yhat_test], axis=1)
d_sub.columns = ['Id', 'SalePrice']
d_sub.head(5)
```

Out[8]:

	Id	SalePrice
0	1461	124430.333281
1	1462	155592.000000
2	1463	177635.099062
3	1464	185549.633437
4	1465	187703.900000

In [9]:
```
# Write to CSV
d_sub.to_csv("./kaggle_submission/baseline_submission.csv", index = False)
```

Kaggle Submission

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/submit

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Make a submission for **To sell, or not to sell**

You have 9 submissions remaining today. This resets 15 hours from now (00: 00 UTC).

Step 1
Upload submission file

Upload Submission File

baseline_submission.csv (22.01 KB) Complete 100% 22.01 KB

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 1459 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

Styling with Markdown supported

H2O Random Forest Baseline

Make Submission

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

3,877 teams · 2 years to go

Overview Data Kernels Discussion Leaderboard Rules Team | My Submissions | Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
baseline_submission.csv	a few seconds ago	0 seconds	0 seconds	0.14421

Complete

Jump to your position on the leaderboard

You can select up to 5 submissions to be used to calculate your final leaderboard score. If 5 submissions are not selected, they will be chosen based on your best submission scores on the public leaderboard.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

30 submissions for To sell, or not to sell — Sort by Most recent

All Successful Selected

Submission and Description	Public Score	Use for Final Score
baseline_submission.csv a few seconds ago by Jo-fai Chow H2O Random Forest Baseline	0.14421	☐

Baseline Model Summary

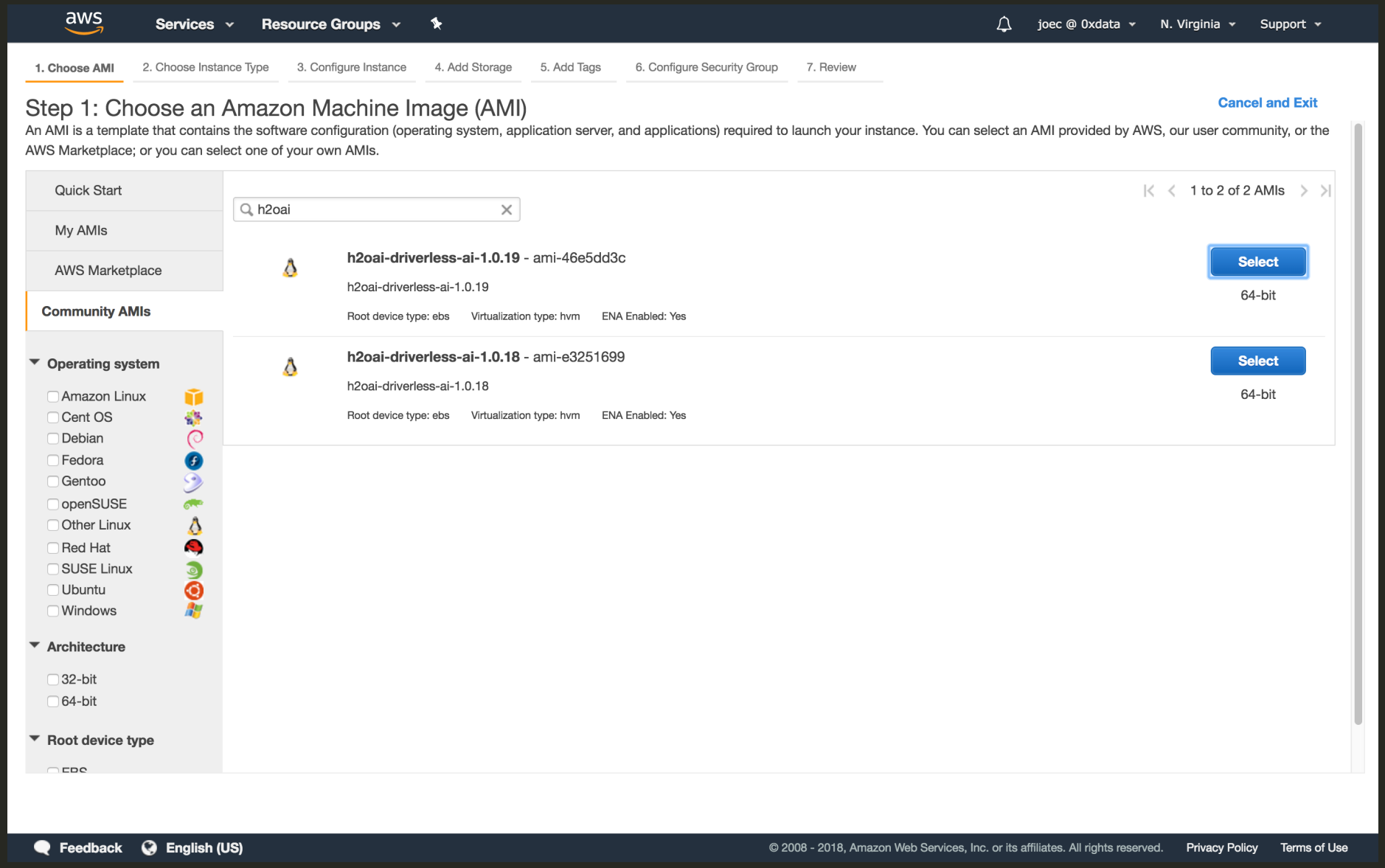
RMSLE (5-Fold CV): 0.13963

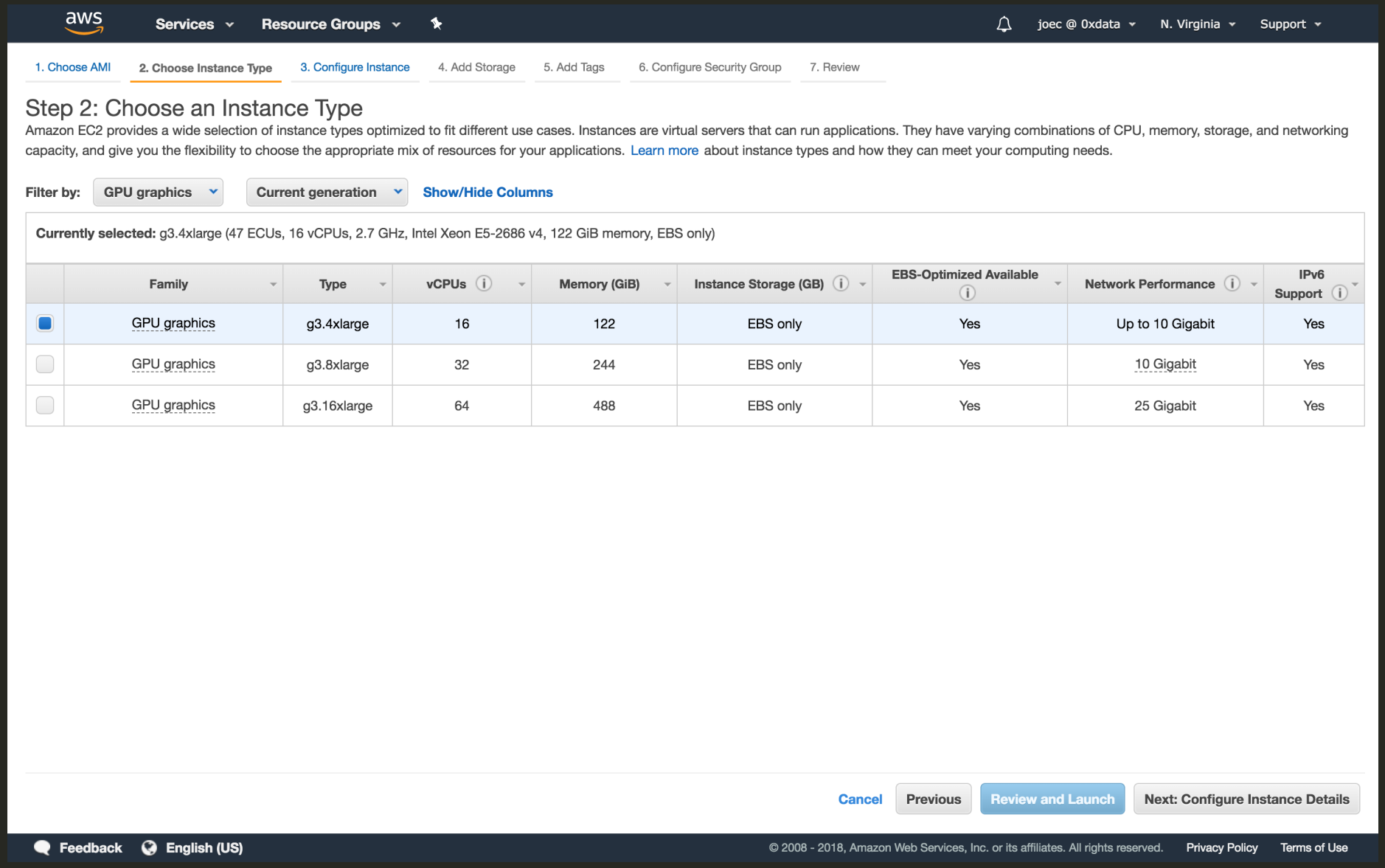
RMSLE (Kaggle Leaderboard): 0.14421

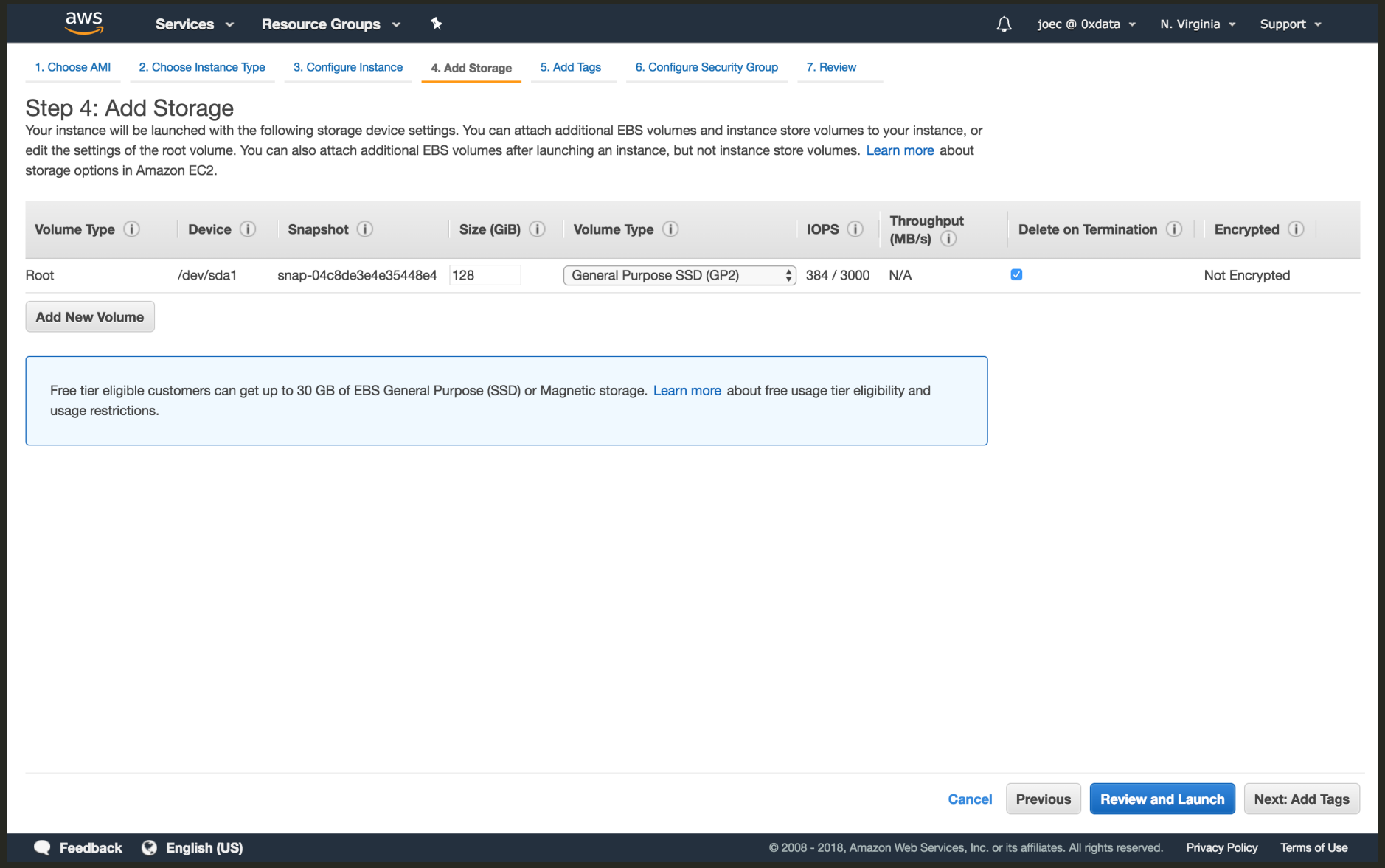
Next: Beat the baseline score using Driverless AI

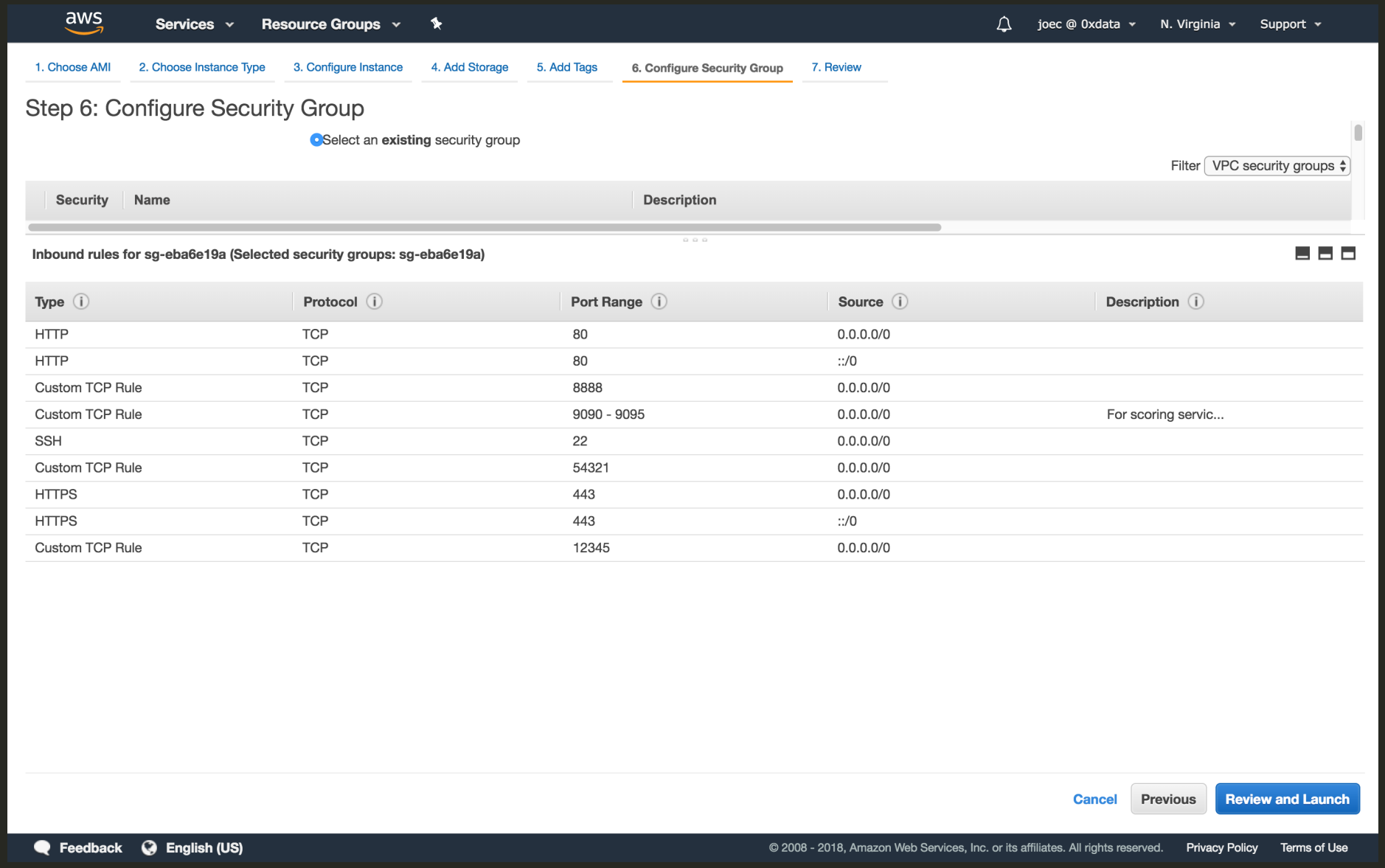
Using Driverless AI

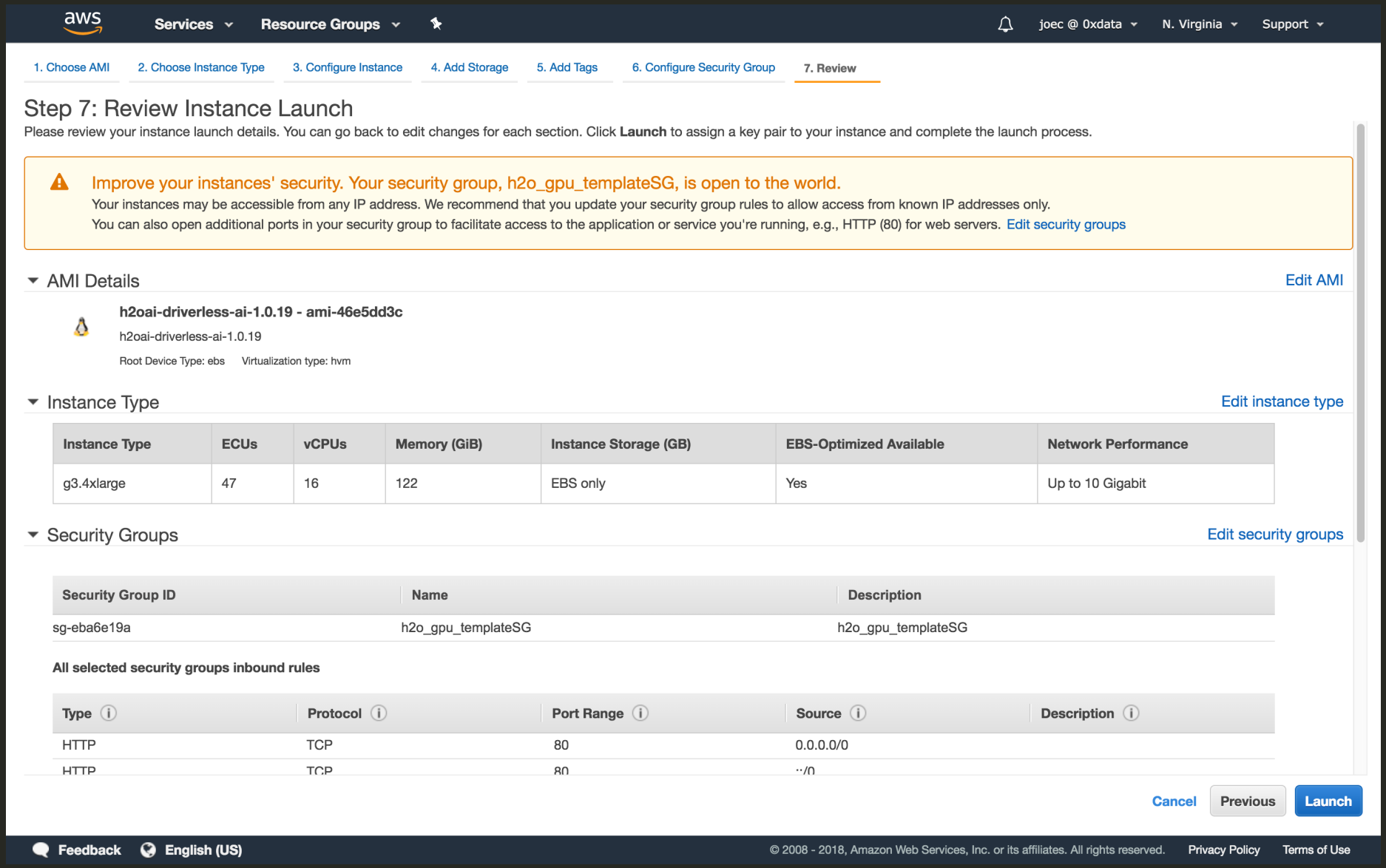
Community AMI on AWS

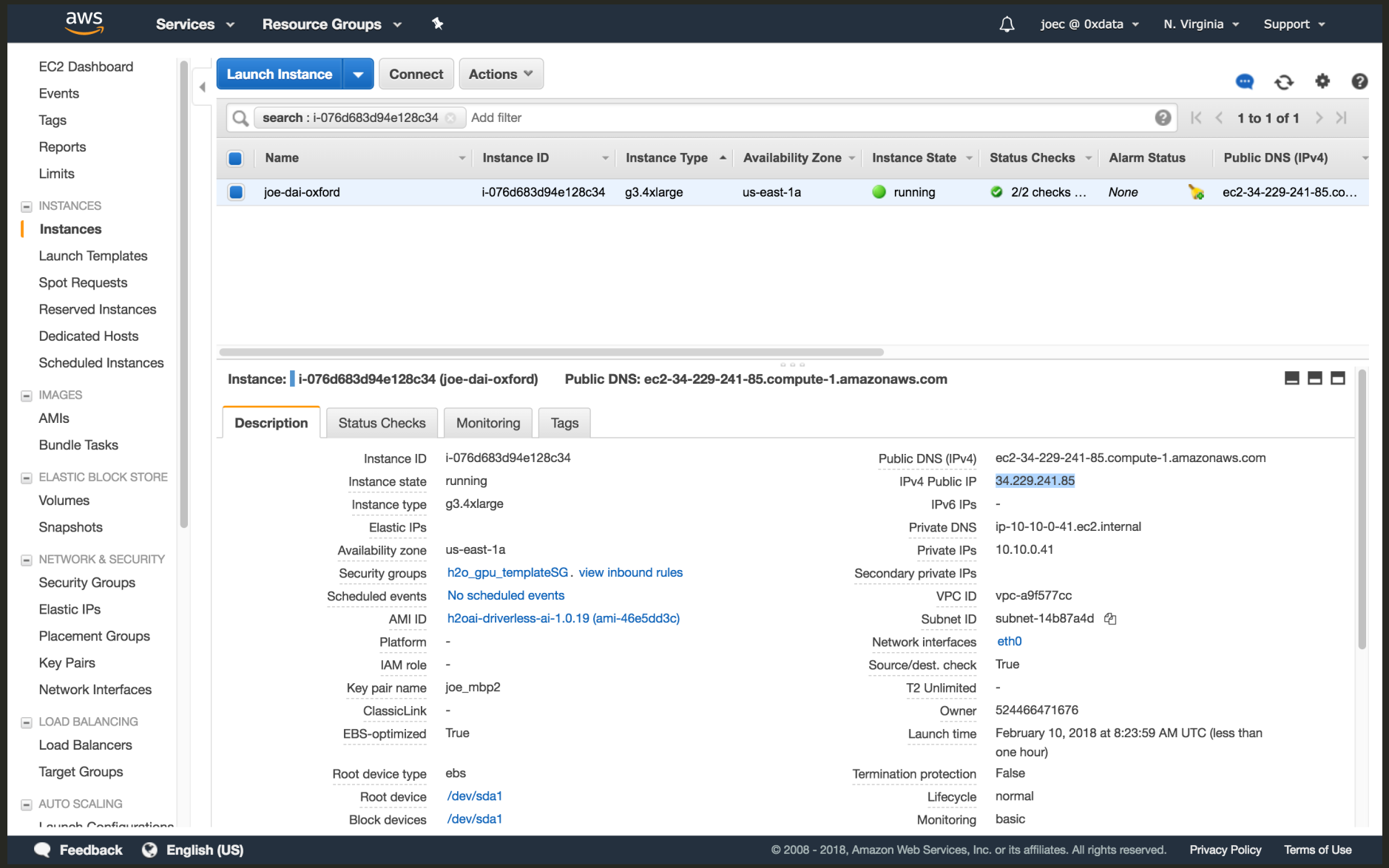

aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 1: Choose an Amazon Machine Image (AMI)
Cancel and Exit
An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. You can select an AMI provided by AWS, our user community, or the AWS Marketplace; or you can select one of your own AMIs.
Quick Start
My AMIs
AWS Marketplace
Community AMIs
Operating system
Amazon Linux
Cent OS
Debian
Fedora
Gentoo
openSUSE
Other Linux
Red Hat
SUSE Linux
Ubuntu
Windows
Architecture
32-bit
64-bit
Root device type
h2oai
1 to 2 of 2 AMIs
h2oai-driverless-ai-1.0.19 - ami-46e5dd3c
h2oai-driverless-ai-1.0.19
Root device type: ebs
Virtualization type: hvm
ENA Enabled: Yes
Select
64-bit
h2oai-driverless-ai-1.0.18 - ami-e3251699
h2oai-driverless-ai-1.0.18
Root device type: ebs
Virtualization type: hvm
ENA Enabled: Yes
Select
64-bit
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 2: Choose an Instance Type
Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can run applications. They have varying combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. Learn more about instance types and how they can meet your computing needs.
Filter by:
GPU graphics
Current generation
Show/Hide Columns
Currently selected: g3.4xlarge (47 ECUs, 16 vCPUs, 2.7 GHz, Intel Xeon E5-2686 v4, 122 GiB memory, EBS only)
Family
Type
vCPUs
Memory (GiB)
Instance Storage (GB)
EBS-Optimized Available
Network Performance
IPv6 Support
GPU graphics g3.4xlarge 16 122 EBS only Yes Up to 10 Gigabit Yes
GPU graphics g3.8xlarge 32 244 EBS only Yes 10 Gigabit Yes
GPU graphics g3.16xlarge 64 488 EBS only Yes 25 Gigabit Yes
Cancel
Previous
Review and Launch
Next: Configure Instance Details
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 4: Add Storage
Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. Learn more about storage options in Amazon EC2.
Volume Type
Device
Snapshot
Size (GiB)
Volume Type
IOPS
Throughput (MB/s)
Delete on Termination
Encrypted
Root
/dev/sda1
snap-04c8de3e4e35448e4
128
General Purpose SSD (GP2)
384 / 3000
N/A
Not Encrypted
Add New Volume
Free tier eligible customers can get up to 30 GB of EBS General Purpose (SSD) or Magnetic storage. Learn more about free usage tier eligibility and usage restrictions.
Cancel
Previous
Review and Launch
Next: Add Tags
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 6: Configure Security Group
Select an existing security group
Filter
VPC security groups
Security
Name
Description
Inbound rules for sg-eba6e19a (Selected security groups: sg-eba6e19a)
Type
Protocol
Port Range
Source
Description
HTTP TCP 80 0.0.0.0/0
HTTP TCP 80 ::/0
Custom TCP Rule TCP 8888 0.0.0.0/0
Custom TCP Rule TCP 9090 - 9095 0.0.0.0/0 For scoring servic...
SSH TCP 22 0.0.0.0/0
Custom TCP Rule TCP 54321 0.0.0.0/0
HTTPS TCP 443 0.0.0.0/0
HTTPS TCP 443 ::/0
Custom TCP Rule TCP 12345 0.0.0.0/0
Cancel
Previous
Review and Launch
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
1. Choose AMI
2. Choose Instance Type
3. Configure Instance
4. Add Storage
5. Add Tags
6. Configure Security Group
7. Review
Step 7: Review Instance Launch
Please review your instance launch details. You can go back to edit changes for each section. Click Launch to assign a key pair to your instance and complete the launch process.
Improve your instances' security. Your security group, h2o_gpu_templateSG, is open to the world.
Your instances may be accessible from any IP address. We recommend that you update your security group rules to allow access from known IP addresses only.
You can also open additional ports in your security group to facilitate access to the application or service you're running, e.g., HTTP (80) for web servers. Edit security groups
AMI Details
Edit AMI
h2oai-driverless-ai-1.0.19 - ami-46e5dd3c
h2oai-driverless-ai-1.0.19
Root Device Type: ebs Virtualization type: hvm
Instance Type
Edit instance type
Instance Type | ECUs | vCPUs | Memory (GiB) | Instance Storage (GB) | EBS-Optimized Available | Network Performance
g3.4xlarge | 47 | 16 | 122 | EBS only | Yes | Up to 10 Gigabit
Security Groups
Edit security groups
Security Group ID | Name | Description
sg-eba6e19a | h2o_gpu_templateSG | h2o_gpu_templateSG
All selected security groups inbound rules
Type | Protocol | Port Range | Source | Description
HTTP | TCP | 80 | 0.0.0.0/0
Cancel
Previous
Launch
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use


aws
Services
Resource Groups
joec @ 0xdata
N. Virginia
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Launch Templates
Spot Requests
Reserved Instances
Dedicated Hosts
Scheduled Instances
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces
LOAD BALANCING
Load Balancers
Target Groups
AUTO SCALING
Launch Instance
Connect
Actions
search : i-076d683d94e128c34
Add filter
1 to 1 of 1
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
Alarm Status
Public DNS (IPv4)
joe-dai-oxford
i-076d683d94e128c34
g3.4xlarge
us-east-1a
running
2/2 checks ...
None
ec2-34-229-241-85.co...
Instance: i-076d683d94e128c34 (joe-dai-oxford)
Public DNS: ec2-34-229-241-85.compute-1.amazonaws.com
Description
Status Checks
Monitoring
Tags
Instance ID
i-076d683d94e128c34
Instance state
running
Instance type
g3.4xlarge
Elastic IPs
Availability zone
us-east-1a
Security groups
h2o_gpu_templateSG. view inbound rules
Scheduled events
No scheduled events
AMI ID
h2oai-driverless-ai-1.0.19 (ami-46e5dd3c)
Platform
-
IAM role
-
Key pair name
joe_mbp2
ClassicLink
-
EBS-optimized
True
Root device type
ebs
Root device
/dev/sda1
Block devices
/dev/sda1
Public DNS (IPv4)
ec2-34-229-241-85.compute-1.amazonaws.com
IPv4 Public IP
34.229.241.85
IPv6 IPs
-
Private DNS
ip-10-10-0-41.ec2.internal
Private IPs
10.10.0.41
Secondary private IPs
VPC ID
vpc-a9f577cc
Subnet ID
subnet-14b87a4d
Network interfaces
eth0
Source/dest. check
True
T2 Unlimited
-
Owner
524466471676
Launch time
February 10, 2018 at 8:23:59 AM UTC (less than one hour)
Termination protection
False
Lifecycle
normal
Monitoring
basic
Feedback
English (US)
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

I will give you the IP address

Username: iot

Password: iot

H2O.ai
1.0.19

License file does not exist (/license/license.sig)

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

ENTER LICENSE

Datasets overview

+ ADD DATASET

No datasets available. Import or drag supported file type from your file browser (mouseover for list of supported types).

© 2017-2018 H2O.ai. All rights reserved.

H2O.ai
1.0.19

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

Datasets overview

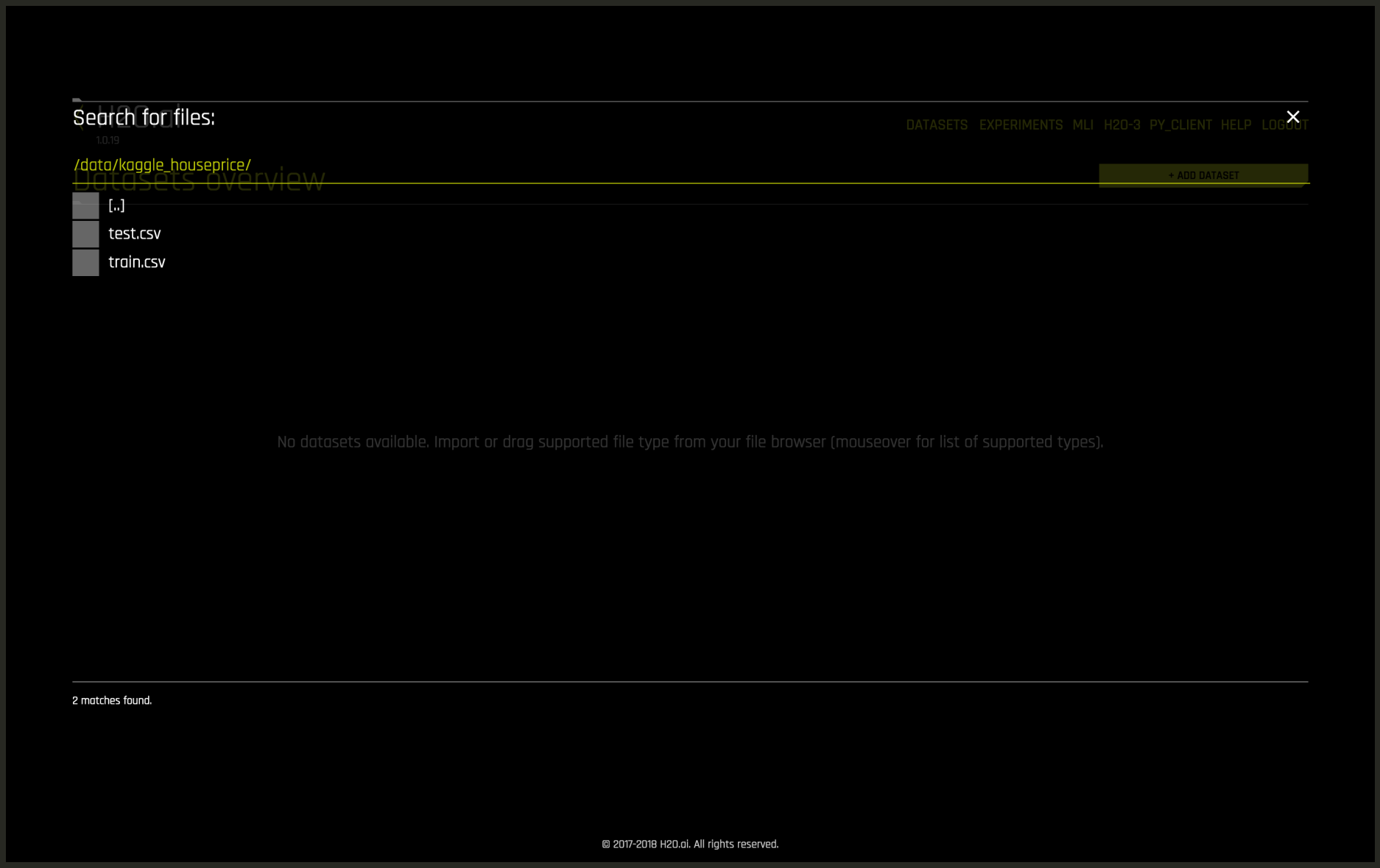
+ ADD DATASET

No datasets available. Import or drag supported file type from your file browser (mouseover for list of supported types).

© 2017-2018 H2O.ai. All rights reserved.

Using Driverless AI

Auto Visualisation


Search for files:
DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT
1.0.19
/data/kaggle_houseprice/
Datasets overview
+ ADD DATASET
[..]
test.csv
train.csv
No datasets available. Import or drag supported file type from your file browser (mouseover for list of supported types).
2 matches found.
© 2017-2018 H2O.ai. All rights reserved.

H2O.ai
1.0.19

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

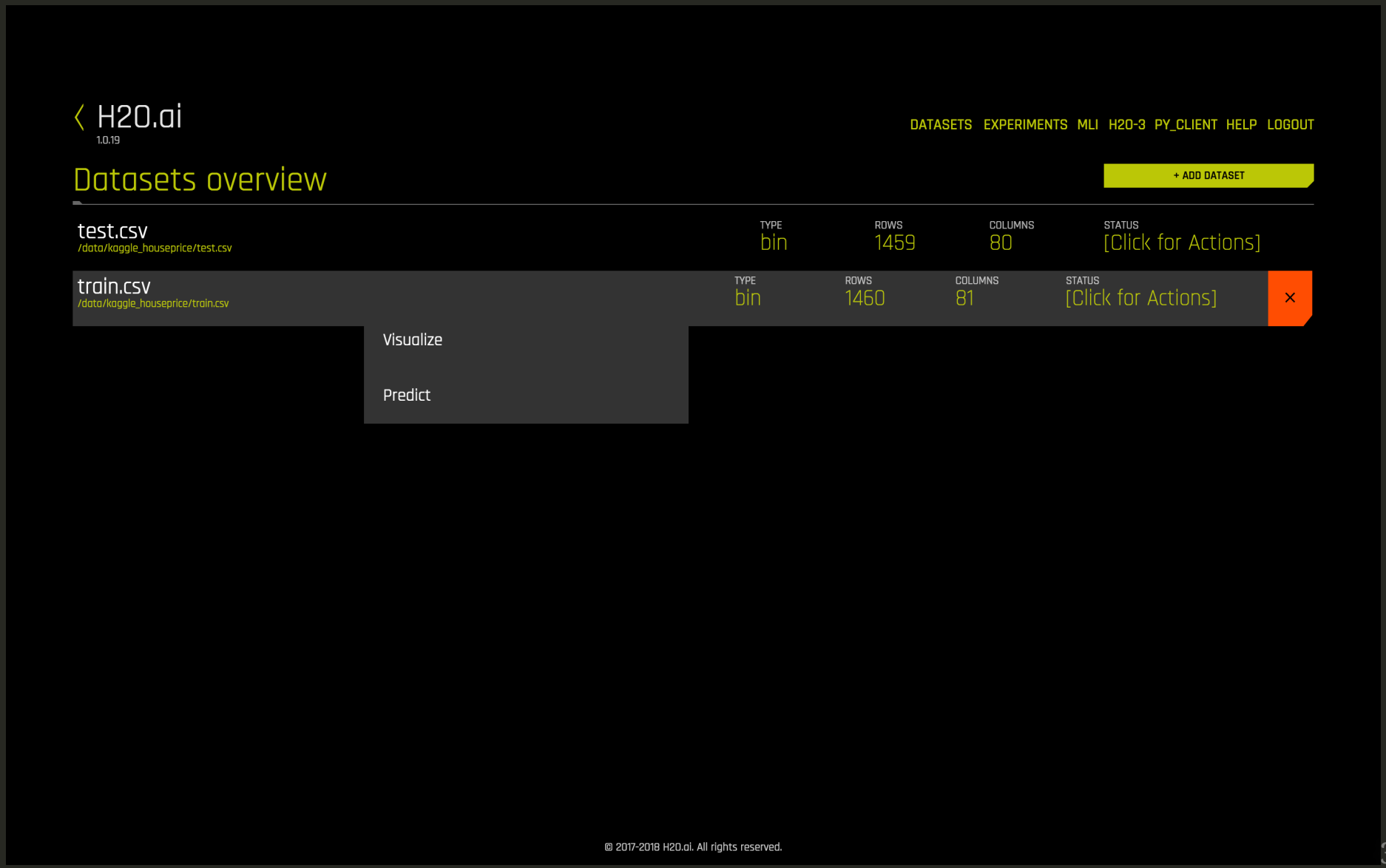
Datasets overview

+ ADD DATASET

	TYPE	ROWS	COLUMNS	STATUS
test.csv /data/kaggle_houseprice/test.csv	bin	1459	80	[Click for Actions]
train.csv /data/kaggle_houseprice/train.csv	bin	1460	81	[Click for Actions]

© 2017-2018 H2O.ai. All rights reserved.

Click on dataset -> "Visualize"

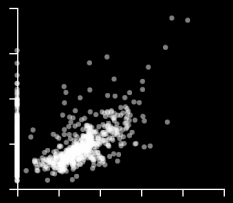


H2O.ai
1.0.19

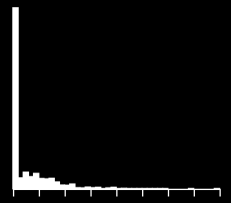
DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

Visualizations for: train.csv

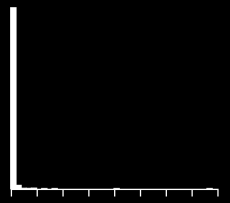
CLUMPY SCATTERPLOTS



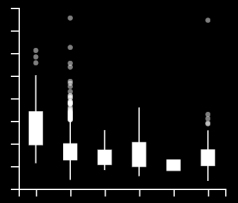
SPIKEY HISTOGRAMS



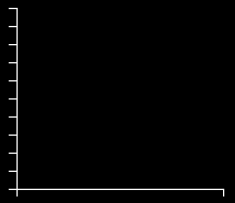
SKEWED HISTOGRAMS



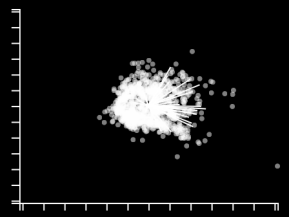
VARYING BOXPLOTS



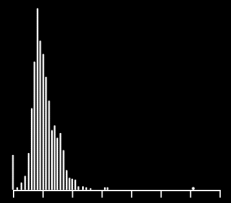
DISPARATE BOXPLOTS



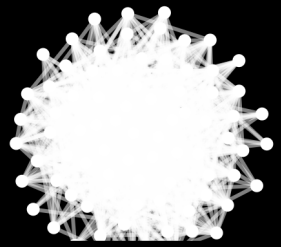
BIPLOTS



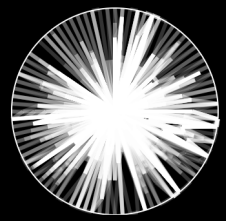
OUTLIERS



CORRELATION GRAPH



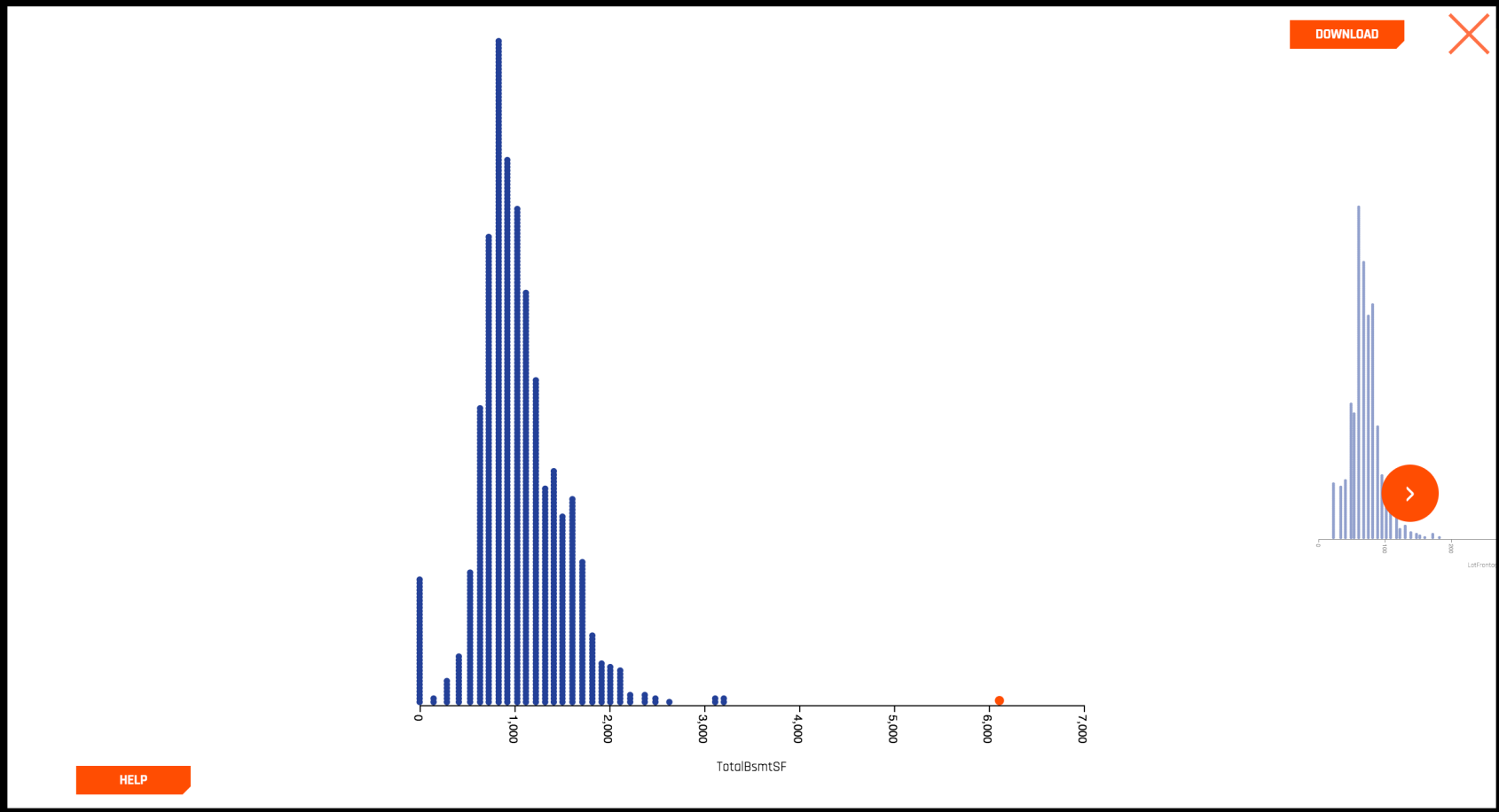
RADAR PLOT



DATA HEATMAP

MISSING HEATMAP

© 2017-2018 H2O.ai. All rights reserved.



© 2017-2018 H2O.ai. All rights reserved.

Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborho...	Condition1	Condi
1299	60	RL	313	63887	Pave	IR3	Bnk		AllPub	Corner	Gtl	Edwards	Feedr	Norm

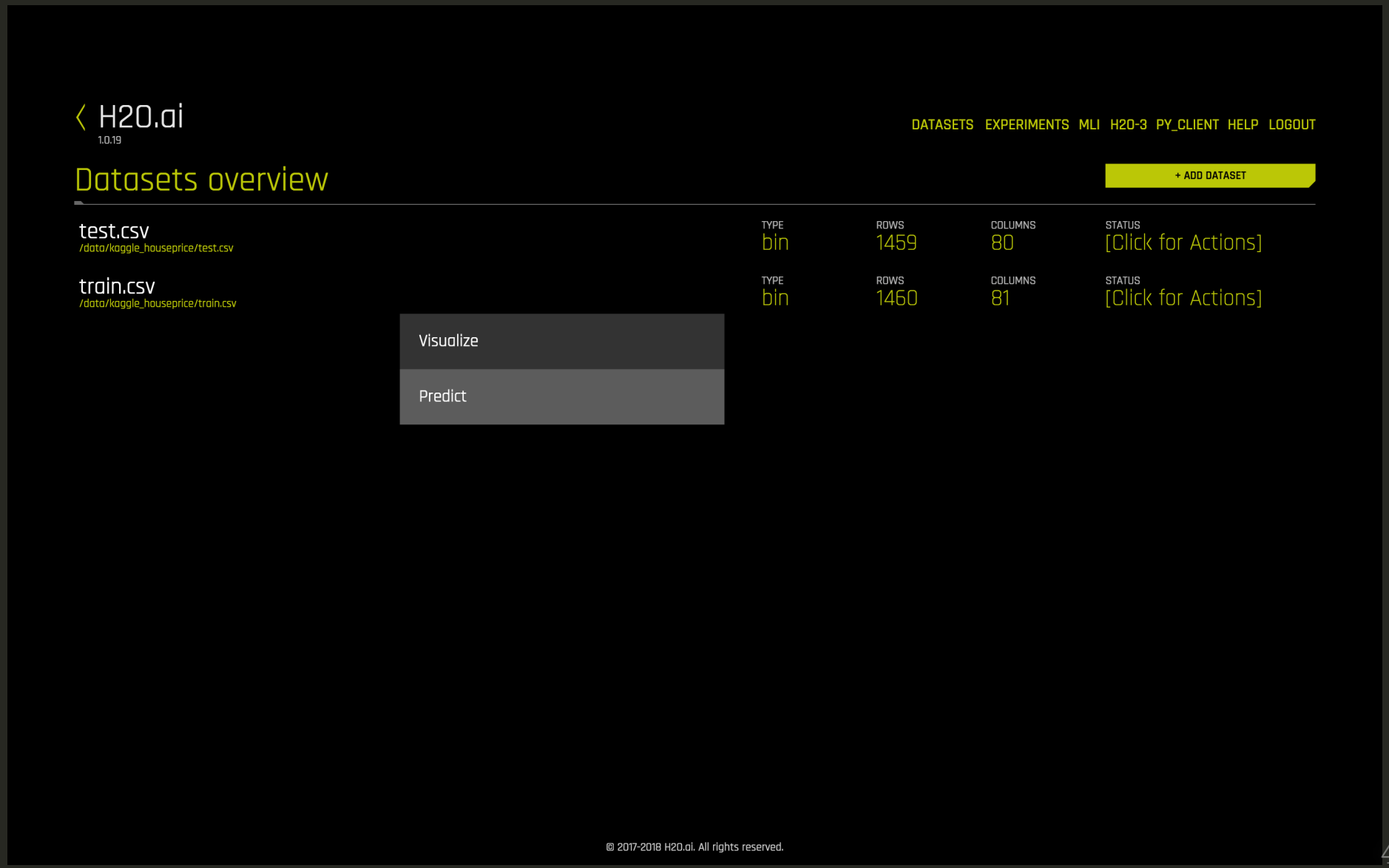
Rows 1-1 of 1 total

© 2017-2018 H2O.ai. All rights reserved.

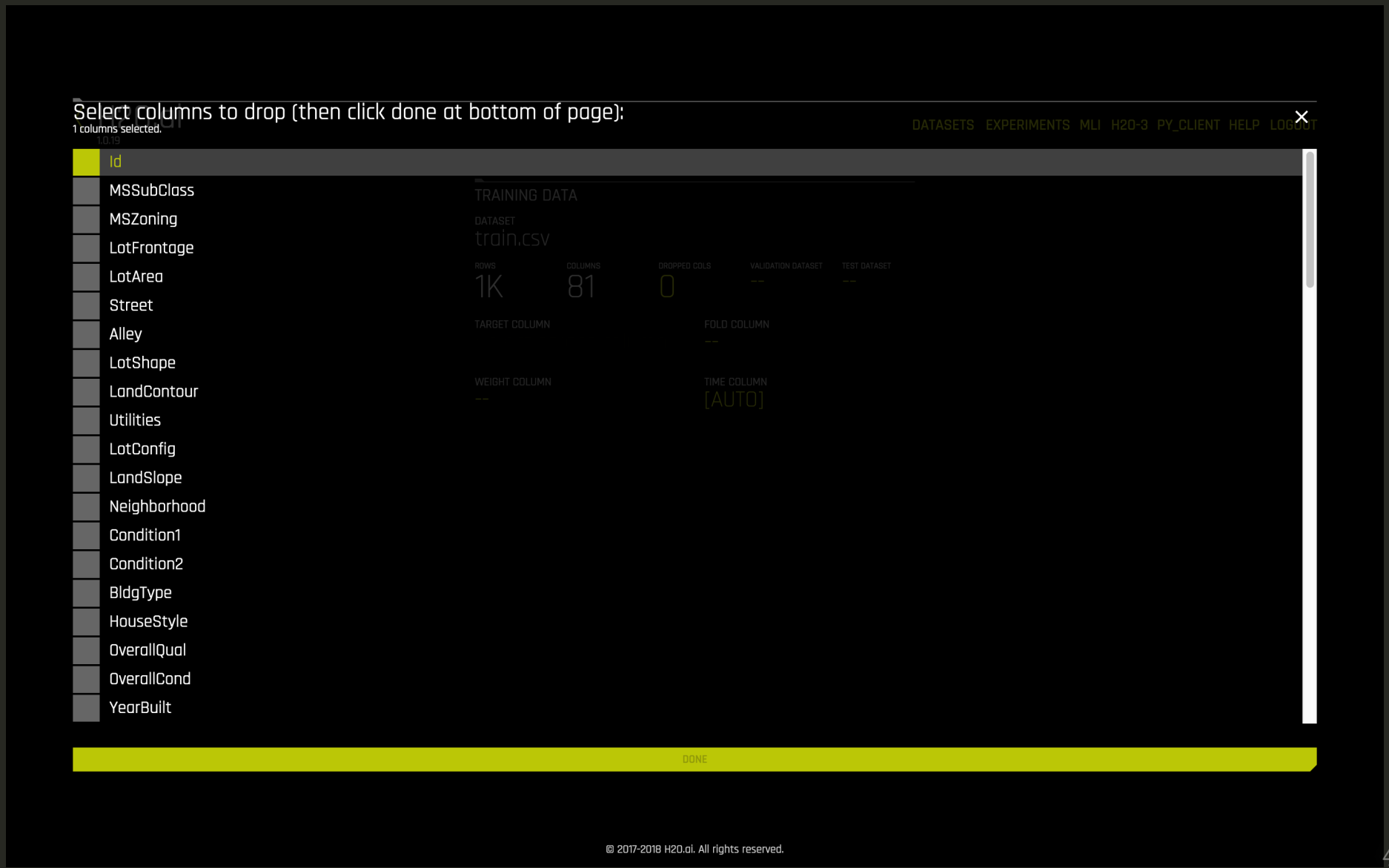
Using Driverless AI

Making Predictions & Beating the Baseline

Click on dataset -> "Predict"



Drop column "id"



© 2017-2018 H2O.ai. All rights reserved.

- Add "Test" dataset
- Select "SalePrice" as target
- Time Column "OFF"
- Settings: Accuracy 5, Time 1, Interpretability 1 (for a quick run)
- Scorer: "RMSLE"
- "Launch Experiment"

H2O.ai Experiment gucitohu

1.0.19

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

TRAINING DATA

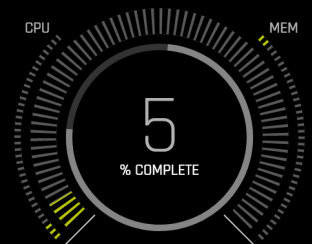
DATASET
train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
1K	81	1	--	Yes (test.csv)

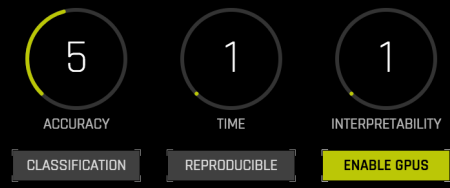
TARGET COLUMN: SalePrice
FOLD COLUMN: --
WEIGHT COLUMN: --
TIME COLUMN: [OFF]

TYPE	COUNT	MEAN	STD DEV
int	1460	180921.196	79442.503

TRAINED 0/27 TARGET TRANSFORM MODELS

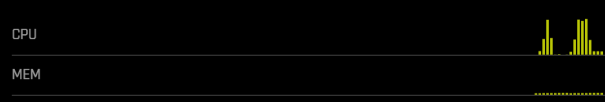


EXPERIMENT SETTINGS



SCORER: GINI, R2, MSE, RMSE, RMSLE, RMSPE, MAE, MAPE, SMAPE

CPU / MEMORY

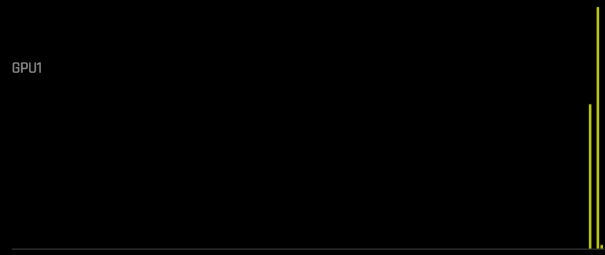


ITERATION SCORES - VALIDATION

VARIABLE IMPORTANCE

93_OverallQual	1.00
51_CVTE:OverallQual.0	0.39
81_GrLivArea	0.17
27_CVTE:GarageCars.0	0.10
49_CVTE:Neighborhood.0	0.09
70_BsmtFinSF1	0.07
97_TotalBsmtSF	0.06
78_GarageArea	0.05
66_1stFlrSF	0.04
96_TotRmsAbvGrd	0.03
67_2ndFlrSF	0.03
25_CVTE:FullBath.0	0.03
84_LotArea	0.02
11_CVTE:BsmtQual.0	0.02

GPU USAGE



© 2017-2018 H2O.ai. All rights reserved.

H2O.ai Experiment gucitohu

1.0.19

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

TRAINING DATA

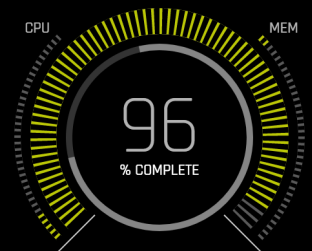
DATASET
train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
1K	81	1	--	Yes (test.csv)

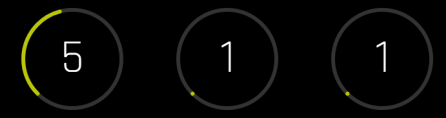
TARGET COLUMN: SalePrice
FOLD COLUMN: --
WEIGHT COLUMN: --
TIME COLUMN: [OFF]

TYPE	COUNT	MEAN	STD DEV
int	1460	180921.196	79442.503

TRAINED 1/8 ENSEMBLE BASE LEARNERS



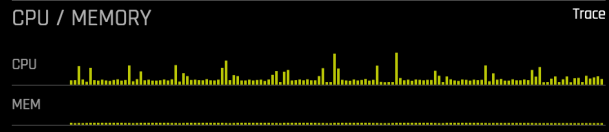
EXPERIMENT SETTINGS



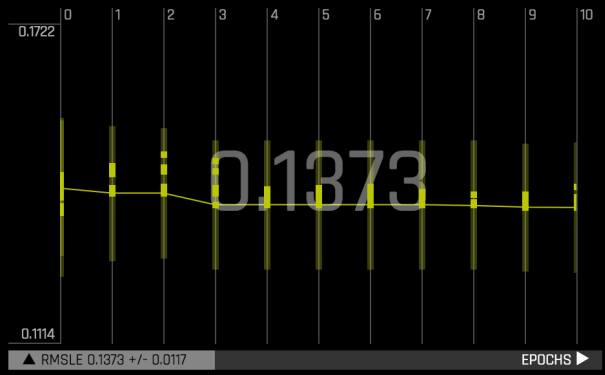
CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

SCORER: GINI, R2, MSE, RMSE, RMSLE, RMSPE, MAE, MAPE, SMAPE

CPU / MEMORY



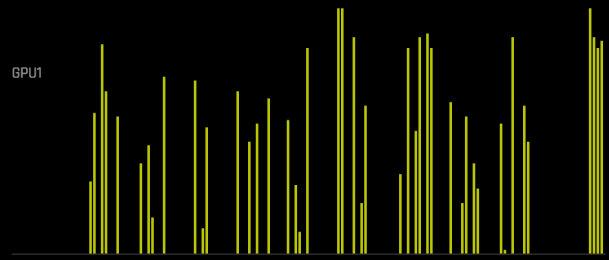
ITERATION SCORES - VALIDATION



VARIABLE IMPORTANCE

110_InteractionMul:OverallQual:YearRemodAdd	1.00
104_TruncSVD:2ndFlrSF:TotalBsmtSF.0	0.46
81_GrLivArea	0.14
49_CVTE:Neighborhood.0	0.10
102_NumCatTE:BsmtQual:FireplaceQu:FullBath:HalfBath:...	0.09
109_TruncSVD:Fireplaces:GarageCars:GrLivArea:HalfBath:...	0.09
114_TruncSVD:2ndFlrSF:TotalBsmtSF.0	0.04
70_BsmtFinSF1	0.03
113_NumCatTE:Fireplaces:Functional:GarageCars:LotArea.0	0.02
84_LotArea	0.02
37_CVTE:KitchenQual.0	0.02
66_1stFlrSF	0.02
12_CVTE:CentralAir.0	0.02
103_CVTE:GarageCars:Street.0	0.02

GPU USAGE



© 2017-2018 H2O.ai. All rights reserved.

H2O.ai Experiment gucitohu

1.0.19

DATASETS EXPERIMENTS MLI H2O-3 PY_CLIENT HELP LOGOUT

TRAINING DATA

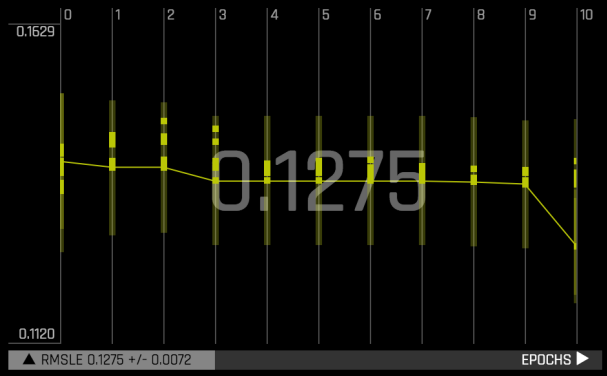
DATASET
train.csv

ROWS	COLUMNS	DROPPED COLS	VALIDATION DATASET	TEST DATASET
1K	81	1	--	Yes (test.csv)

TARGET COLUMN: SalePrice
FOLD COLUMN: --
WEIGHT COLUMN: --
TIME COLUMN: [OFF]

TYPE	COUNT	MEAN	STD DEV
int	1460	180921.196	79442.503

ITERATION SCORES - VALIDATION



STATUS: COMPLETE

- INTERPRET THIS MODEL
- SCORE ON ANOTHER DATASET
- TRANSFORM ANOTHER DATASET...
- DOWNLOAD (HOLDOUT) TRAINING PREDICTIONS
- DOWNLOAD TEST PREDICTIONS
- DOWNLOAD LOGS
- DOWNLOAD SCORING PIPELINE

SUMMARY

VARIABLE IMPORTANCE

110_InteractionMul:OverallQual:YearRemodAdd	1.00
104_TruncSVD:2ndFlrSF:TotalBsmtSF.0	0.46
81_GrLivArea	0.14
49_CVTE:Neighborhood.0	0.10
102_NumCatTE:BsmtQual:FireplaceQu:FullBath:HalfBath:...	0.09
109_TruncSVD:Fireplaces:GarageCars:GrLivArea:HalfBath:...	0.09
114_TruncSVD:2ndFlrSF:TotalBsmtSF.0	0.04
70_BsmtFinSF1	0.03
113_NumCatTE:Fireplaces:Functional:GarageCars:LotArea.0	0.02
84_LotArea	0.02
37_CVTE:KitchenQual.0	0.02
66_1stFlrSF	0.02
12_CVTE:CentralAir.0	0.02
103_CVTE:GarageCars:Street.0	0.02

EXPERIMENT SETTINGS

ACCURACY	TIME	INTERPRETABILITY
5	1	1

CLASSIFICATION | REPRODUCIBLE | ENABLE GPUS

SCORER: GINI, R2, MSE, RMSE, RMSLE, RMSPE, MAE, MAPE, SMAPE

CPU / MEMORY

Trace

EXPERIMENT SUMMARY

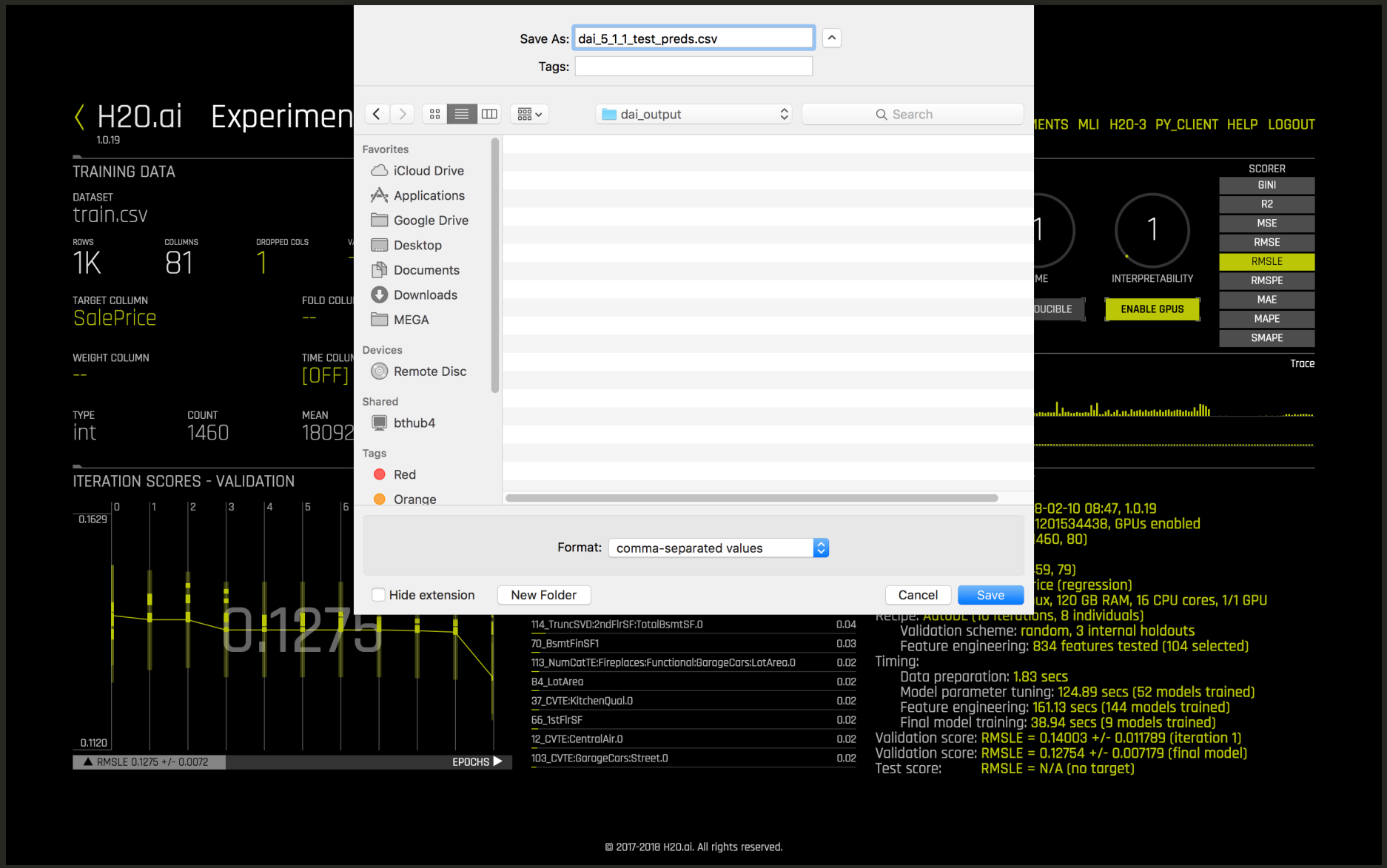
Experiment: gucitohu, 2018-02-10 08:47, 1.0.19
- Settings: 5/1/1, seed=1201534438, GPUs enabled
- Train data: train.csv (1460, 80)
- Validation data: N/A
- Test data: test.csv (1459, 79)
- Target column: SalePrice (regression)

System specs: Docker/Linux, 120 GB RAM, 16 CPU cores, 1/1 GPU
Recipe: AutoDL (10 iterations, 8 individuals)
- Validation scheme: random, 3 internal holdouts
- Feature engineering: 834 features tested (104 selected)

Timing:
- Data preparation: 1.83 secs
- Model parameter tuning: 124.89 secs (52 models trained)
- Feature engineering: 161.13 secs (144 models trained)
- Final model training: 38.94 secs (9 models trained)

Validation score: RMSLE = 0.14003 +/- 0.011789 (iteration 1)
Validation score: RMSLE = 0.12754 +/- 0.007179 (final model)
Test score: RMSLE = N/A (no target)

© 2017-2018 H2O.ai. All rights reserved.


H20.ai Experimen
1.0.19
TRAINING DATA
DATASET
train.csv
ROWS
1K
COLUMNS
81
DROPPED COLS
1
TARGET COLUMN
SalePrice
WEIGHT COLUMN
--
TIME COLU
[OFF]
TYPE
int
COUNT
1460
MEAN
18092
ITERATION SCORES - VALIDATION
0.1629
0.1275
0.1120
RMSLE 0.1275 +/- 0.0072
EPOCHS
Save As: dai_5_1_1_test_preds.csv
Tags:
dai_output
Search
Favorites
iCloud Drive
Applications
Google Drive
Desktop
Documents
Downloads
MEGA
Devices
Remote Disc
Shared
bthub4
Tags
Red
Orange
Format: comma-separated values
Hide extension
New Folder
Cancel
Save
MENTS MLI H2O-3 PY_CLIENT HELP LOGOUT
SCORER
GINI
R2
MSE
RMSE
RMSLE
RMSPE
MAE
MAPE
SMAPE
INTERPRETABILITY
ENABLE GPUS
Trace
114_TruncSVD:2ndFlrSF:TotalBsmtSF.0 0.04
70_BsmtFinSF1 0.03
113_NumCatTE:Fireplaces:Functional:GarageCars:LotArea.0 0.02
84_LotArea 0.02
37_CVTE:KitchenQual.0 0.02
66_1stFlrSF 0.02
12_CVTE:CentralAir.0 0.02
103_CVTE:GarageCars:Street.0 0.02
8-02-10 08:47, 1.0.19
1201534438, GPUs enabled
460, 80)
59, 79)
ice (regression)
ux, 120 GB RAM, 16 CPU cores, 1/1 GPU
Validation scheme: random, 3 internal holdouts
Feature engineering: 834 features tested (104 selected)
Timing:
Data preparation: 1.83 secs
Model parameter tuning: 124.89 secs (52 models trained)
Feature engineering: 161.13 secs (144 models trained)
Final model training: 38.94 secs (9 models trained)
Validation score: RMSLE = 0.14003 +/- 0.011789 (iteration 1)
Validation score: RMSLE = 0.12754 +/- 0.007179 (final model)
Test score: RMSLE = N/A (no target)
© 2017-2018 H2O.ai. All rights reserved.

Driverless AI

Reformat "test_preds.csv" and create Kaggle submission

https://github.com/woobe/dai_oxford/blob/master/create_dai_sub.ipynb

This repository Search Pull requests Issues Marketplace Explore

woobe / dai_oxford

Unwatch 1 Star 0 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master ▾ dai_oxford / create_dai_sub.ipynb Find file Copy path

woobe Fixed typo 3ac9024 24 seconds ago

1 contributor

133 lines (132 sloc) | 3.13 KB Raw Blame History

In [1]:
```
import pandas as pd
```

In [2]:
```
# Create Kaggle Submission
d_sub = pd.read_csv("./data/sample_submission.csv")
dai_pred = pd.read_csv("./dai_output/dai_5_1_1_test_preds.csv") # <----- change this!!!

d_sub = pd.concat([d_sub['Id'].reset_index(drop=True), dai_pred], axis=1)
d_sub.columns = ['Id', 'SalePrice']
d_sub.head(5)
```

Out[2]:

	Id	SalePrice
0	1461	127100.231248
1	1462	155623.797813
2	1463	178983.809033
3	1464	187495.775536
4	1465	183158.616124

In [3]:
```
# Write to CSV
d_sub.to_csv("./kaggle_submission/dai_5_1_1_submission.csv", index = False) # <------ change this!!!
```

House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

3,877 teams · 2 years to go

Overview Data Kernels Discussion Leaderboard Rules Team | My Submissions | Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
dai_5_1_1_submission.csv	just now	0 seconds	0 seconds	0.12709

Complete

Jump to your position on the leaderboard

You can select up to 5 submissions to be used to calculate your final leaderboard score. If 5 submissions are not selected, they will be chosen based on your best submission scores on the public leaderboard.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

30 submissions for To sell, or not to sell — Sort by Most recent

All Successful Selected

Submission and Description	Public Score	Use for Final Score
dai_5_1_1_submission.csv just now by Jo-fai Chow DAI 5-1-1 Quick Run	0.12709	☐

Summary

Baseline Model:

- RMSLE (5-Fold CV): 0.13963
- RMSLE (Kaggle Leaderboard): 0.14421

Driverless AI Quick Run (5-1-1):

- RMSLE (3-Fold CV): 0.12754
- RMSLE (Kaggle Leaderboard): 0.12709

(Now it is your turn to improve this score)

Beyond this Exercise

- Your own feature engineering + Driverless AI
- Kaggle Kernels / Discussion
- Try different algorithms

Thanks!

Contact: Slack / joe@h2o.ai

Slides created via the R package **xaringan**.