notebook_regression_supermarket_sales

Assistance

	Routine	*Description*
	`importFiles`	Import file(s) into H_2O
	`importSqlTable`	Import SQL table into H_2O
	`getFrames`	Get a list of frames in H_2O
	`splitFrame`	Split a frame into two or more frames
	`mergeFrames`	Merge two frames into one
	`getModels`	Get a list of models in H_2O
	`getGrids`	Get a list of grid search results in H_2O
	`getPredictions`	Get a list of predictions in H_2O
	`getJobs`	Get a list of jobs running in H_2O
	`runAutoML`	Automatically train and tune many models
	`buildModel`	Build a model
	`importModel`	Import a saved model
	`predict`	Make a prediction

Import Files

Search: Enter a file or directory path and press the Enter key

Selected Files: (No files selected)

Actions: Import

1 / 1 files imported.

Files /Users/apple/Downloads/dataset_regression_supermarket_sales.csv

Actions Parse these files...

Setup Parse

PARSE CONFIGURATION

Sources nfs://Users/apple/Downloads/dataset_regression_supermarket_sales.csv

ID dataset_regression_supermarket_sales.hex

Parser CSV

Separator ,: '044'

Escape Character 0

Column Headers
- Auto
- First row contains column names
- First row contains data

Options
- Enable single quotes as a field quotation character
- Delete on done

EDIT COLUMN NAMES AND TYPES

Search by column name...

1	Invoice ID	String	750-67-8428	226-31-3081	631-41-3108	123-19-1176	373-73-7910	699-14-3026	355-53-5
2	Branch	Enum	A	C	A	A	A	C	A
3	City	Enum	Yangon	Naypyitaw	Yangon	Yangon	Yangon	Naypyitaw	Yangon
4	Customer ty	Enum	Member	Normal	Normal	Member	Normal	Normal	Member
5	Gender	Enum	Female	Female	Male	Male	Male	Male	Female
6	Product lir	Enum	Health and beauty	Electronic accessories	Home and lifestyle	Health and beauty	Sports and travel	Electronic accessories	Electroı accesso
7	Unit price	Numeric	74.69	15.28	46.33	58.22	86.31	85.39	68.84
8	Quantity	Numeric	7	5	7	8	7	7	6
9	Tax	Numeric	26.1415	3.82	16.2155	23.288	30.2085	29.8865	20.652

10	Total	Numeric	548.9715	80.22	340.5255	489.048	634.3785	627.6165	433.692
11	Date	Time	1/5/2019	3/8/2019	3/3/2019	1/27/2019	2/8/2019	3/25/2019	2/25/20
12	Time	Enum	13:08	10:29	13:23	20:33	10:37	18:30	14:36
13	Payment	Enum	Ewallet	Cash	Credit card	Ewallet	Ewallet	Ewallet	Ewallet
14	CostOfGood:	Numeric	522.83	76.4	324.31	465.76	604.17	597.73	413.04
15	gross marg	Numeric	4.761904762	4.761904762	4.761904762	4.761904762	4.761904762	4.761904762	4.76190

Previous page Next page

Parse

Job

Run Time 00:00:00.209

Remaining Time 00:00:00.0

Type Frame

Key dataset_regression_supermarket_sales.hex

Description Parse

Status DONE

Progress 100%

Done.

Actions View

dataset_regression_supermarket_sales.hex

Actions: View Data | Split | Build Model | Run AutoML | Predict | Download | Export | Delete

Rows	*Columns*	*Compressed Size*
1000	18	247KB

COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinali
Invoice ID	string	0	0	0	0	·	·	·	·	
Branch	enum	0	340	0	0	0	2.0	·	·	
City	enum	0	332	0	0	0	2.0	·	·	
Customer type	enum	0	501	0	0	0	1.0	0.4990	0.5002	
Gender	enum	0	501	0	0	0	1.0	0.4990	0.5002	
Product line	enum	0	170	0	0	0	5.0	·	·	
Unit price	real	0	0	0	0	10.0800	99.9600	55.6721	26.4946	
Quantity	int	0	0	0	0	1.0	10.0	5.5100	2.9234	
Tax	real	0	0	0	0	0.5085	49.6500	15.3794	11.7088	
Total	real	0	0	0	0	10.6785	1042.6500	322.9667	245.8853	
Date	time	0	0	0	0	1546300800000.0	1553904000000.0	1550102745600.0	2170551497.8377	
Ti		0	2	0	0	0	595.0			5

Previous 20 Columns | Next 20 Columns

▸ CHUNK COMPRESSION SUMMARY

▸ FRAME DISTRIBUTION SUMMARY

dataset_regression_supermarket_sales.hex

▸ DATA

CS

Split Frame

Frame: dataset_regression_supermarket_sales.hex

Splits:	*Ratio*	*Key*
	0.80	dataset_regression_supermarket_sales.hex_0.80
	0.20	dataset_regression_supermarket_sales.hex_0.20

Add a new split

Seed: 43

Create

Split Frames

Type	*Key*	*Ratio*
	dataset_regression_supermarket_sales.hex_0.80	0.8
	dataset_regression_supermarket_sales.hex_0.20	0.19999999999999996

Run AutoML

PARAMETERS

project_name	supermarket_sales_model_1	Optional project name used to group models from multiple AutoML runs into a single Leaderboard; derived from the training data name if not specified.
distribution	AUTO	Distribution function used by algorithms that support it; other algorithms use their defaults.

*training_frame**	dataset_regression_supermarket_sales.hex_0.80	ID of the training data frame.
*response_column**	RatingDecimal	Response column
validation_frame	(Choose...)	ID of the validation data frame (used for early stopping in grid searches and for early stopping of the AutoML process itself).
blending_frame	(Choose...)	ID of the H2OFrame used to train the the metalearning algorithm in Stacked Ensembles (instead of relying on cross-validated predicted values). When provided, it is also recommended to disable cross validation by setting `nfolds=0` and to provide a leaderboard frame for scoring purposes.
leaderboard_frame	(Choose...)	ID of the leaderboard data frame (used to score models and rank them on the AutoML Leaderboard).

ADVANCED

nfolds	-1	Number of folds for k-fold cross-validation (defaults to -1 (AUTO), otherwise it must be >=2 or use 0 to disable). Disabling prevents Stacked Ensembles from being built.
balance_classes	☐	Balance training data class counts via over/under-sampling (for imbalanced data).
custom_metric_func		Reference to custom evaluation function, format: `language:keyName=funcName`

exclude_algos — Search...

- ☐ GLM
- ☐ DRF
- ☐ GBM
- ☐ DeepLearning
- ☐ StackedEnsemble
- ☐ XGBoost

☑ All ☐ None

A list of algorithms to skip during the model-building phase.

exploitation_ratio — -1

The budget ratio (between 0 and 1) dedicated to the exploitation (vs exploration) phase.

monotone_constraints — Choose... ↔ +

A mapping representing monotonic constraints. Use +1 to enforce an increasing constraint and -1 to specify a decreasing constraint.

fold_column — (Choose...)

Fold column (contains fold IDs) in the training frame. These assignments are used to create the folds for cross-validation of the models.

weights_column — (Choose...)

Weights column in the training frame, which specifies the row weights used in model training.

ignored_columns Search...

Showing page 1 of 2. 15 ignored.

☐ Invoice ID	STRING
☑ Branch	ENUM(3)
☑ City	ENUM(3)
☑ Customer type	ENUM(2)
☑ Gender	ENUM(2)
☑ Product line	ENUM(6)
☑ Unit price	REAL
☑ Quantity	INT
☑ Tax	REAL
☑ Total	REAL

☑ All ☐ None ← Previous 10 → Next 10

Only show columns with more than 0 % missing values.

Names of columns to ignore in the training frame when building models.

sort_metric AUTO

Metric used to sort leaderboard

seed -1

Seed for random number generator; set to a value other than -1 for reproducibility.

max_models 0

Maximum number of models to build (optional). Always set this parameter to ensure AutoML reproducibility: all models are then trained until convergence and none is constrained by a time budget.

max_runtime_secs	0	This argument specifies the maximum time that the AutoML process will run for. If both max_runtime_secs and max_models are specified, then the AutoML run will stop as soon as it hits either of these limits. If neither max_runtime_secs nor max_models are specified, then max_runtime_secs defaults to 3600 seconds (1 hour).
max_runtime_secs_per_model	0	Maximum time to spend on each individual model (optional). Note that models constrained by a time budget are not guaranteed reproducible.
stopping_rounds	3	Early stopping based on convergence of stopping_metric. Stop if simple moving average of length k of the stopping_metric does not improve for k:=stopping_rounds scoring events (0 to disable)
stopping_metric	AUTO	Metric to use for early stopping (AUTO: logloss for classification, deviance for regression)
stopping_tolerance	-1	Relative tolerance for metric-based stopping criterion (stop if relative improvement is not at least this much)

EXPERT

keep_cross_validation_predictions	☑	Whether to keep the predictions of the cross-validation predictions. This needs to be set to TRUE if running the same AutoML object for repeated runs because CV predictions are required to build additional Stacked Ensemble models in AutoML.
keep_cross_validation_models	☑	Whether to keep the cross-validated models. Keeping cross-validation models may consume significantly more memory in the H2O cluster.
keep_cross_validation_fold_assignment	☐	Whether to keep cross-validation assignments.
export_checkpoints_dir		Path to a directory where every generated model will be stored.

Build Models

Job

Run Time 00:45:24.311

Remaining Time 00:00:00.0

Type AutoML

Key supermarket_sales_model_1@@RatingDecimal

Description AutoML build

Status DONE

Progress 100%

Done.

Actions View

Leaderboard

Monitor Live

MODELS

models sorted in order of rmse, best first

	model_id	rmse	mse	mae
0	DeepLearning_grid_1_AutoML_1_20260206_144622_model_41	1.2933761330594369	1.6728218215677821	1.0607
1	DeepLearning_grid_1_AutoML_1_20260206_144622_model_20	1.2935516674994527	1.6732759164906146	1.0607
2	DeepLearning_grid_1_AutoML_1_20260206_144622_model_1	1.293568706754643	1.67331999909488	1.0618
3	DeepLearning_grid_1_AutoML_1_20260206_144622_model_24	1.2936325203159658	1.6734850976190379	1.0607
4	DeepLearning_grid_1_AutoML_1_20260206_144622_model_4	1.294320325685819	1.6752651054834442	1.0646
5	DeepLearning_grid_1_AutoML_1_20260206_144622_model_3	1.2944472350020122	1.6755936442043549	1.0638
6	DeepLearning_grid_1_AutoML_1_20260206_144622_model_12	1.2944786740938352	1.6756750376837335	1.0626
7	DeepLearning_grid_1_AutoML_1_20260206_144622_model_31	1.2945166820625917	1.6757734401383415	1.0612
8	DeepLearning_grid_1_AutoML_1_20260206_144622_model_35	1.2946477653894282	1.6761128364278397	1.0613
9	DeepLearning_grid_1_AutoML_1_20260206_144622_model_26	1.2950352605583126	1.677116326089337	1.0634
10	DeepLearning_grid_1_AutoML_1_20260206_144622_model_22	1.295062321121066	1.6771864155874827	1.0593
11	DeepLearning_grid_1_AutoML_1_20260206_144622_model_23	1.2950750977505614	1.677219508813626	1.0607
12	DeepLearning_grid_1_AutoML_1_20260206_144622_model_34	1.2952275241421538	1.6776143392954137	1.0645
13	DeepLearning_grid_3_AutoML_1_20260206_144622_model_6	1.2952632109602342	1.677706785667016	1.0651
14	DeepLearning_grid_3_AutoML_1_20260206_144622_model_9	1.295439749264555	1.678164143974613	1.0644
15	DeepLearning_grid_1_AutoML_1_20260206_144622_model_11	1.2954468517386986	1.6781825456797057	1.0677
16	DeepLearning_grid_1_AutoML_1_20260206_144622_model_37	1.295457485537006	1.678210096833862	1.0625
17	DeepLearning_grid_2_AutoML_1_20260206_144622_model_2	1.295505145911152	1.678333583082275	1.0669
18	DeepLearning_grid_1_AutoML_1_20260206_144622_model_39	1.2955595853276884	1.678474639134452	1.0639
19	StackedEnsemble_BestOfFamily_6_AutoML_1_20260206_144622	1.2957117129891444	1.678868843177263	1.0632
20	DeepLearning_grid_1_AutoML_1_20260206_144622_model_21	1.2957311494352273	1.6789192116167353	1.0627
21	DeepLearning_grid_1_AutoML_1_20260206_144622_model_6	1.2957481213203523	1.6789631939052228	1.0630
22	DeepLearning_grid_1_AutoML_1_20260206_144622_model_44	1.29579420423666	1.6790826197333188	1.0628

EVENT LOG

Actions taken and discoveries made by AutoML

44	14:46:23.790	DEBUG	ModelTraining	leaderboard Leaderboard_supermarket_sales_model_1@@RatingDecimal. Training time: model=0s, total=0s	
45	14:46:23.792	DEBUG	ModelTraining	Time assigned for XGBoost_2_AutoML_1_20260206_144622: 553.6275625s	
46	14:46:23.792	INFO	ModelTraining	AutoML: starting XGBoost_2_AutoML_1_20260206_144622 model training	st
47	14:46:23.792	WARN	ModelTraining	_train param, Dropping bad and constant columns: [gross margin percentage, Invoice ID]	
48	14:46:23.792	DEBUG	ModelTraining	XGBoost_2_AutoML_1_20260206_144622 [XGBoost def_1] started	
49	14:46:23.962	DEBUG	ModelTraining	XGBoost_2_AutoML_1_20260206_144622 [XGBoost def_1] complete	
50	14:46:23.962	DEBUG	ModelTraining	Adding model XGBoost_2_AutoML_1_20260206_144622 to leaderboard Leaderboard_supermarket_sales_model_1@@RatingDecimal. Training time: model=0s, total=0s	
51	14:46:23.966	DEBUG	ModelTraining	Time assigned for DRF_1_AutoML_1_20260206_144622: 654.2555s	
52	14:46:23.966	INFO	ModelTraining	AutoML: starting DRF_1_AutoML_1_20260206_144622 model training	st
53	14:46:23.966	WARN	ModelTraining	_train param, Dropping bad and constant columns: [gross margin percentage, Invoice ID]	
54	14:46:23.966	DEBUG	ModelTraining	DRF_1_AutoML_1_20260206_144622 [DRF def_1] started	
55	14:46:24.254	DEBUG	ModelTraining	DRF_1_AutoML_1_20260206_144622 [DRF def_1] complete	
56	14:46:24.255	DEBUG	ModelTraining	Adding model DRF_1_AutoML_1_20260206_144622 to leaderboard Leaderboard_supermarket_sales_model_1@@RatingDecimal. Training time: model=0s, total=0s	
57	14:46:24.256	DEBUG	ModelTraining	Time assigned for GBM_2_AutoML_1_20260206_144622: 799.581125s	
58	14:46:24.256	INFO	ModelTraining	AutoML: starting GBM_2_AutoML_1_20260206_144622 model training	st
59	14:46:24.257	WARN	ModelTraining	_train param, Dropping bad and constant columns: [gross margin percentage, Invoice ID]	
60	14:46:24.257	DEBUG	ModelTraining	GBM_2_AutoML_1_20260206_144622 [GBM def_2] started	
61	14:46:24.487	DEBUG	ModelTraining	GBM_2_AutoML_1_20260206_144622 [GBM def_2] complete	