Final Flow

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: C:\Users\Nilkanth Jadhav\Desktop\YesBank\Yes_Bank_Training.csv

Search Results: (All files added)

Selected Files: 1 file selected: Clear All

✖ C:\Users\Nilkanth Jadhav\Desktop\YesBank\Yes_Bank_Training.csv

Actions: Import

1 / 1 files imported.

Files C:\Users\Nilkanth Jadhav\Desktop\YesBank\Yes_Bank_Training.csv

Actions Parse these files...

Setup Parse

PARSE CONFIGURATION

Sources nfs:\C:\Users\Nilkanth Jadhav\Desktop\YesBank\Yes_Bank_Training.csv

ID Yes_Bank_Training.hex

Parser CSV

Separator ,: '044'

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	serial_numb	Numeric	1	2	3	4	5	6	7	8	9
2	age_in_year	Numeric	58	44	33	47	33	35	28	42	58
3	job_descrip	Enum	management	technician	entrepreneur	blue-collar	unknown	management	management	entrepreneur	re
4	marital_sta	Enum	married	single	married	married	single	married	single	divorced	ma
5	education_d	Enum	tertiary	secondary	secondary	unknown	unknown	tertiary	tertiary	tertiary	pr
6	has_default	Enum	no	no	no	no	no	no	no	yes	no
7	balance_in_	Numeric	2143	29	2	1506	1	231	447	2	12
8	housing_sta	Enum	yes	yes	yes	yes	no	yes	yes	yes	ye

9	previous_lo	Enum	no	no	yes	no	no	no	yes	no	no
10	phone_type	Enum	unknown	unknown	unknown	unknown	unknown	unknown	unknown	unknown	un
11	date	Numeric	5	5	5	5	5	5	5	5	5
12	month_of_ye	Enum	may	may	may	may	may	may	may	may	ma
13	call_durati	Numeric	261	151	76	92	198	139	217	380	50
14	campaign_co	Numeric	1	1	1	1	1	1	1	1	1
15	days_passed	Numeric	-1	-1	-1	-1	-1	-1	-1	-1	-1

Previous page Next page

Parse

Job

Run Time 00:00:01.653
Remaining Time 00:00:00.0
Type Frame
Key Yes_Bank_Training.hex
Description Parse
Status DONE
Progress 100%

Done.

Actions View

Yes_Bank_Training.hex

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

Rows	*Columns*	*Compressed Size*
31649	18	566KB

COLUMN SUMMARIES

label	*type*	*Missing*	*Zeros*	*+Inf*	*-Inf*	*min*	*max*	*mean*	*sigma*	*cardinality*	*Actions*
serial_number	int	0	0	0	0	1.0	31649.0	15825.0	9136.4237	·	Convert to enum
age_in_years	int	0	0	0	0	19.0	94.0	41.1118	9.5977	·	Convert to enum
job_description	enum	0	3354	0	0	0	11.0	·	·	12	Convert to numeric
marital_status	enum	0	3799	0	0	0	2.0	·	·	3	Convert to numeric
education_details	enum	0	4977	0	0	0	3.0	·	·	4	Convert to numeric
has_default	enum	0	30925	0	0	0	1.0	0.0229	0.1495	2	Convert to numeric
balance_in_account	int	0	2717	0	0	-8019.0	98417.0	1293.3825	2961.1851	·	Convert to enum
housing_status	enum	0	14419	0	0	0	1.0	0.5444	0.4980	2	Convert to numeric
previous_loan	enum	0	25847	0	0	0	1.0	0.1833	0.3869	2	Convert to numeric
phone_type	enum	0	17181	0	0	0	2.0	·	·	3	Convert to numeric
date	enum	0	90	0	0	0	30.0	·	·	31	Convert to numeric
month_of_year	enum	0	190	0	0	0	10.0	·	·	11	Convert to numeric
call_duration	int	0	2	0	0	0	4918.0	252.4082	262.3450	·	Convert to enum
campaign_contacts	int	0	0	0	0	1.0	63.0	3.0624	3.5102	·	Convert to enum
days_passed	int	0	0	0	0	-1.0	335.0	10.8199	48.0868	·	Convert to enum
previous_contact	int	0	29591	0	0	0	275.0	0.2085	1.9463	·	Convert to enum
poutcome_of_campaign	enum	0	1439	0	0	0	3.0	·	·	4	Convert to numeric
outcome	enum	0	29809	0	0	0	1.0	0.0581	0.2340	2	Convert to numeric

Previous 20 Columns Next 20 Columns

CHUNK COMPRESSION SUMMARY

FRAME DISTRIBUTION SUMMARY

Build a Model

Select an algorithm: Gradient Boosting Machine

PARAMETERS GRID?

model_id	gbm-43c4eaef-6aeb-4a7a-b6b3-8121ad087d25	Destination id for this model; auto-generated if not specified.
training_frame	Yes_Bank_Training.hex	Id of the training data frame.

validation_frame (Choose...) — Id of the validation data frame.

nfolds 0 — Number of folds for K-fold cross-validation (0 to disable or >= 2).

response_column outcome — Response variable column.

ignored_columns Search...

Showing page 1 of 1. 10 ignored.

previous_loan	ENUM(2)
phone_type	ENUM(3)
date	ENUM(31)
month_of_year	ENUM(11)
call_duration	INT
campaign_contacts	INT
days_passed	INT
previous_contact	INT
poutcome_of_campaign	ENUM(4)
outcome	ENUM(2)

All | None | Previous 100 | Next 100

Only show columns with more than 0 % missing values.

ignore_const_cols — Ignore constant columns.

ntrees 10000 — Number of trees.

max_depth 6 — Maximum tree depth.

min_rows 10 — Fewest allowed (weighted) observations in a leaf.

nbins 10 — For numerical columns (real/int), build a histogram of (at least) this many bins, then split at the best point

seed -1 — Seed for pseudo random number generator (if applicable)

learn_rate 0.001 — Learning rate (from 0.0 to 1.0)

sample_rate 1 — Row sample rate per tree (from 0.0 to 1.0)

col_sample_rate 1 — Column sample rate (from 0.0 to 1.0)

ADVANCED GRID?

score_each_iteration — Whether to score during each iteration of model training.

score_tree_interval 0 — Score the model after every so many trees. Disabled if set to 0.

fold_column (Choose...) — Column with cross-validation fold index assignment per observation.

offset_column (Choose...) — Offset column. This will be added to the combination of columns before applying the link function.

weights_column (Choose...) — Column with observation weights. Giving some observation a weight of zero is equivalent to excluding it from the dataset; giving an observation a relative weight of 2 is equivalent to repeating that row twice. Negative weights are not allowed. Note: Weights are per-row observation weights and do not increase the size of the data frame. This is typically the number of times a row is repeated, but non-integer values are supported as well. During training, rows with higher weights matter more, due to the larger loss function pre-factor.

balance_classes — Balance training data class counts via over/under-sampling (for imbalanced data).

nbins_top_level 1024 — For numerical columns (real/int), build a histogram of (at most) this many bins at the root level, then decrease by factor of two per level

nbins_cats 1024 — For categorical columns (factors), build a histogram of this many bins, then split at the best point. Higher values can lead to more overfitting.

r2_stopping	1.7976931348623157e+308	r2_stopping is no longer supported and will be ignored if set - please use stopping_rounds, stopping_metric and stopping_tolerance instead. Previous version of H2O would stop making trees when the R^2 metric equals or exceeds this
stopping_rounds	0	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	0.001	Relative tolerance for metric-based stopping criterion (stop if relative improvement is not at least this much)
max_runtime_secs	0	Maximum allowed runtime in seconds for model training. Use 0 to disable.
learn_rate_annealing	1	Scale the learning rate by this factor after each tree (e.g., 0.99 or 0.999)
distribution	AUTO	Distribution function
huber_alpha	0.9	Desired quantile for Huber/M-regression (threshold between quadratic and linear loss, must be between 0 and 1).
checkpoint		Model checkpoint to resume training with.
col_sample_rate_per_tree	1	Column sample rate per tree (from 0.0 to 1.0)
min_split_improvement	0.00001	Minimum relative improvement in squared error reduction for a split to happen
histogram_type	AUTO	What type of histogram to use for finding optimal split points
categorical_encoding	AUTO	Encoding scheme for categorical features
custom_metric_func		Reference to custom evaluation function, format: `language:keyName=funcName`

EXPERT GRID?

class_sampling_factors		Desired over/under-sampling ratios per class (in lexicographic order). If not specified, sampling factors will be automatically computed to obtain class balance during training. Requires balance_classes.
max_after_balance_size	5	Maximum relative size of the training data after balancing class counts (can be less than 1.0). Requires balance_classes.
build_tree_one_node	☐	Run on one node only; no network overhead but fewer cpus used. Suitable for small datasets.
sample_rate_per_class		A list of row sample rates per class (relative fraction for each class, from 0.0 to 1.0), for each tree
col_sample_rate_change_per_level	1	Relative change of the column sampling rate for every level (must be > 0.0 and <= 2.0)
max_abs_leafnode_pred	1.7976931348623157e+308	Maximum absolute value of a leaf node prediction
pred_noise_bandwidth	0	Bandwidth (sigma) of Gaussian multiplicative noise ~N(1,sigma) for tree node predictions
calibrate_model	☑	Use Platt Scaling to calculate calibrated class probabilities. Calibration can provide more accurate estimates of class probabilities.
calibration_frame	Yes_Bank_Training.hex	Calibration frame for Platt Scaling

Build Model

Job

Run Time	00:05:13.491
Remaining Time	00:00:00.0
Type	Grid
Key	grid-09a4f829-e354-4e93-8812-334e8b45b782

Description GBM Grid Search
Status DONE
Progress 100%
Done.
Actions View

Grid Search

Inspect Grid Summary | Inspect Scoring History | Inspect Models | Delete selected

	Key	*Actions*
	grid-09a4f829-e354-4e93-8812-334e8b45b782_model_1	Predict... Inspect

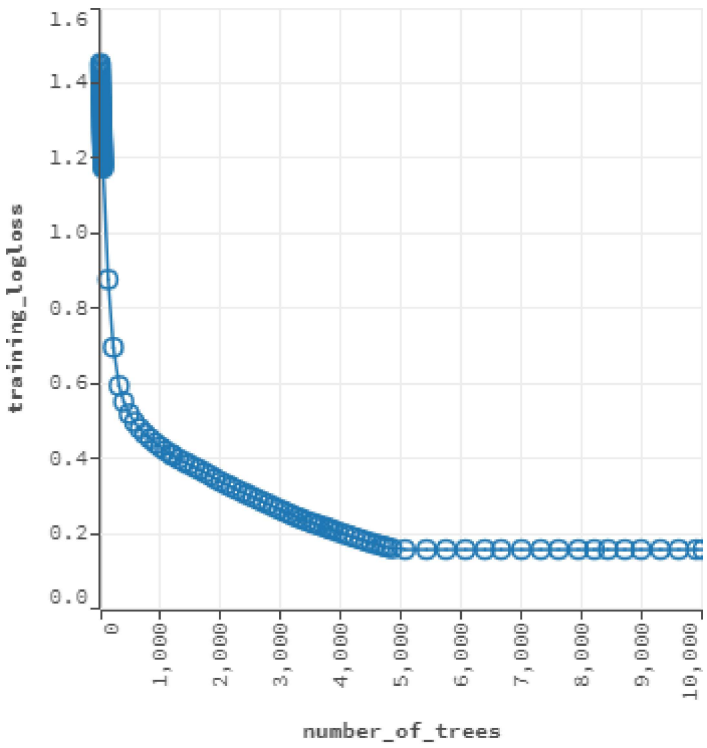
CS

Model

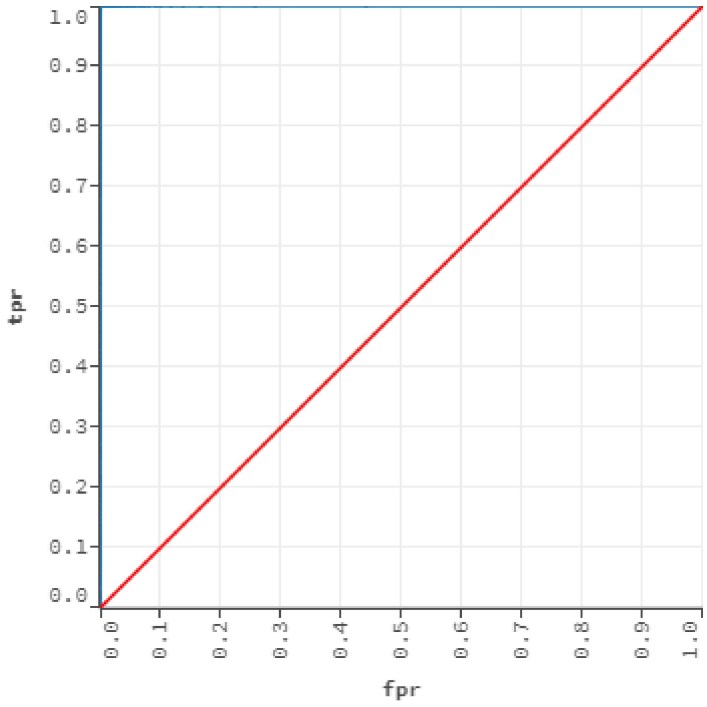
Model ID: grid-09a4f829-e354-4e93-8812-334e8b45b782_model_1
Algorithm: Gradient Boosting Machine
Actions: Stop | Predict... | Download POJO | Download Model Deployment Package (MOJO) | Export | Inspect | Delete | Download Gen Model

▸ MODEL PARAMETERS

▾ SCORING HISTORY - LOGLOSS

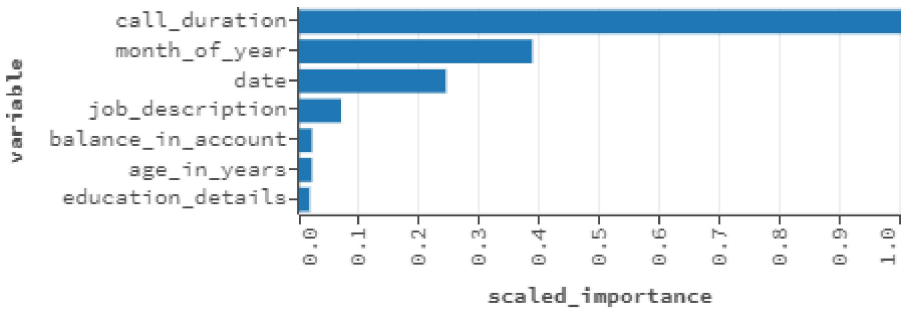


▾ ROC CURVE - TRAINING METRICS , AUC = 0.999965



Threshold: Choose...
Criterion: Choose...

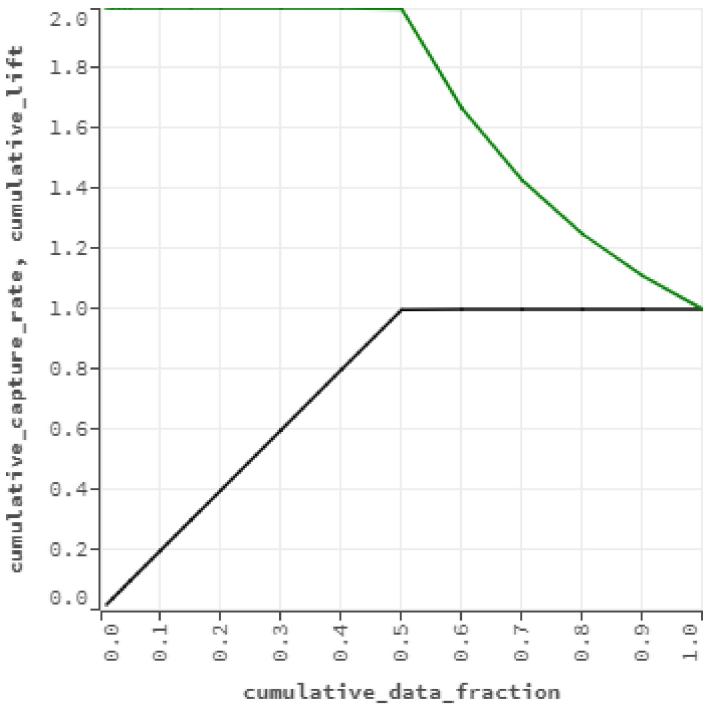
▾ VARIABLE IMPORTANCES



▾ TRAINING METRICS - CONFUSION MATRIX ROW LABELS: ACTUAL CLASS; COLUMN LABELS: PREDICTED CLASS

	no	yes	Error	Rate	Recall
no	29760	49	**0.0016**	**49 / 29,809**	**1.0**
yes	0	29819	**0**	**0 / 29,819**	**1.0**
Total	**29760**	**29868**	**0.0008**	**49 / 59,628**	
Precision	**1.0**	**1.0**			

▾ TRAINING METRICS - GAINS/LIFT TABLE



▸ OUTPUT

▸ OUTPUT - MODEL SUMMARY

▸ OUTPUT - SCORING HISTORY

▸ OUTPUT - TRAINING_METRICS

▸ DOMAIN

▸ OUTPUT - TRAINING_METRICS - METRICS FOR THRESHOLDS (BINOMIAL METRICS AS A FUNCTION OF CLASSIFICATION THRESHOLDS)

▸ OUTPUT - TRAINING_METRICS - MAXIMUM METRICS (MAXIMUM METRICS AT THEIR RESPECTIVE THRESHOLDS)

▸ OUTPUT - TRAINING_METRICS - GAINS/LIFT TABLE (AVG RESPONSE RATE: 50.01 %, AVG SCORE: 37.80 %)

▸ OUTPUT - VARIABLE IMPORTANCES

▾ PREVIEW POJO

</> Preview POJO

Predict

Name: prediction-fe6d9668-57

Model: grid-09a4f829-e354-4e93-8812-334e8b45b782_model_1

Frame: Yes_Bank_Test.hex ▾

Compute Leaf Node Assignment: ☐

Actions: Predict

Prediction

▾ PREDICTION

prediction_frame prediction-fe6d9668-57ae-4d69-a72d-eb6635c7b37b

Combine predictions with frame

prediction-fe6d9668-57ae-4d69-a72d-eb6635c7b37b

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

Rows	*Columns*	*Compressed Size*
13562	3	224KB

COLUMN SUMMARIES

label	*type*	*Missing*	*Zeros*	*+Inf*	*-Inf*	*min*	*max*	*mean*	*sigma*	*cardinality*	*Actions*
predict	enum	0	12446	0	0	0	1.0	0.0823	0.2748	2	Convert to numeric
no	real	0	0	0	0	0.0006	1.0	0.9415	0.1779	·	·
yes	real	0	0	0	0	0.0	0.9994	0.0585	0.1779	·	·

Previous 20 Columns | Next 20 Columns

CHUNK COMPRESSION SUMMARY

FRAME DISTRIBUTION SUMMARY