Driverless AI Experiment: `bosemoke`

Generated on: Tue Dec 4 12:40:30 2018

Generated by: h2oai

Table of Contents

Experiment Overview

Driverless AI built 1 XGBoostModel to predict `X15.9.F.12_qo_boepd` given 42 original features from the input dataset `make2p2i_train.csv`. This regression experiment completed in 21 minutes and 58 seconds (0:21:58), using 1 of the 42 original features, and 1 of the 1,499 engineered features.

Performance

Dataset	RMSE
Internal Validation	248.136
Test Data	Test Data not Provided

Driverless Settings

Dial Settings	Description	Setting Value	Range of Possible Values
Accuracy	Controls sophistication of the model	10	1-10
Time	Controls duration of the experiment	4	1-10
Interpretability	Controls complexity of the model	8	1-10

System Specifications

System	System Memory	CPUs	GPUs
Docker/Linux	60	4	1

Versions

Driverless AI Version
1.4.1

Data Overview

This section provides information on the datasets used for the experiment.

data	file path	number of rows	number of columns
training	./tmp/gicagowu/make2p2i_train.csv.1543925877.14501.bin	112	64
validation	Not provided		
testing	Not provided		

Training Data

The training data consists of only numeric columns

The summary of the columns is shown below:

Numeric Columns

name	data_type	min	mean	max	std	unique	freq of mode
mth_yr	time [%Y-%m-%d]					112	1
X15.9.F.1.C_Gp	real	1,597,936.650	18,013,752.648	26,440,918.600	6,849,774.337	25	1
X15.9.F.1.C_Np	real	11,142.470	122,162.972	177,709.330	45,770.511	25	1
X15.9.F.1.C_On.Stream	real	27.500	399.341	745.000	203.975	25	1
X15.9.F.1.C_qgp_boepd	real	0.005	0.205	0.663	0.160	25	1
X15.9.F.1.C_qo_boepd	real	24.407	1,473.248	5,052.686	1,193.944	25	1
X15.9.F.1.C_qwp_boepd	real	0.000	1,725.418	3,582.094	1,009.800	25	1
X15.9.F.1.C_Wp	real	0.000	86,895.110	207,302.390	67,581.768	25	1
X15.9.F.11_Gp	real	590,504.520	92,933,348.527	174,310,520.610	56,126,672.933	39	1
X15.9.F.11_Np	real	3,923.080	614,834.562	1,147,849.100	371,939.494	39	1
X15.9.F.11_On.Stream	real	112.915	666.826	745.000	117.335	37	3
X15.9.F.11_qgp_boepd	real	0.112	0.864	1.560	0.322	39	1
X15.9.F.11_qo_boepd	real	796.006	6,084.765	11,438.524	2,400.973	39	1
X15.9.F.11_qwp_boepd	real	0.000	5,782.492	17,807.231	5,750.955	39	1
X15.9.F.11_Wp	real	0.000	276,797.811	1,090,806.270	330,717.996	39	1

name	data_type	min	mean	max	std	unique	freq of mode
X15.9.F.12_Gp	real	7,068,009.290	501,324,801.630	667,542,278.020	184,309,440.530	102	2
X15.9.F.12_Np	real	49,091.060	3,463,405.659	4,579,609.550	1,268,977.607	102	2
X15.9.F.12_On.Stream	real	0.000	626.962	745.000	160.114	97	5
X15.9.F.12_qgp_boepd	real	0.000	1.242	4.574	1.297	103	2
X15.9.F.12_qo_boepd	real	0.000	9,112.608	33,771.146	9,822.115	103	2
X15.9.F.12_qwp_boepd	real	0.000	13,599.593	31,876.605	10,957.073	101	4
X15.9.F.12_Wp	real	412.610	3,014,343.672	6,833,320.370	2,554,695.106	100	3
X15.9.F.14_Gp	real	0.000	401,240,014.295	578,009,542.020	183,444,123.849	98	5
X15.9.F.14_Np	real	0.000	2,763,004.687	3,942,233.390	1,245,373.581	98	5
X15.9.F.14_On.Stream	real	0.000	603.593	745.000	206.469	87	7
X15.9.F.14_qgp_boepd	real	0.000	1.075	3.857	0.974	98	7
X15.9.F.14_qo_boepd	real	0.000	7,839.428	28,978.042	7,356.290	98	7
X15.9.F.14_qwp_boepd	real	0.000	14,167.005	28,086.446	8,280.464	98	7
X15.9.F.14_Wp	real	0.000	2,878,935.544	7,121,249.740	2,387,204.430	98	5
X15.9.F.15.D_Gp	real	820,968.960	13,386,982.967	22,505,350.950	6,575,539.349	30	3
X15.9.F.15.D_Np	real	5,674.630	89,185.569	148,518.560	43,378.960	30	3
X15.9.F.15.D_On.Stream	real	0.000	540.146	744.000	230.555	31	3
X15.9.F.15.D_qgp_boepd	real	0.000	0.133	0.251	0.064	31	3
X15.9.F.15.D_qo_boepd	real	0.000	935.318	1,774.364	455.101	31	3
X15.9.F.15.D_qwp_boepd	real	0.000	329.844	1,323.376	460.716	22	12
X15.9.F.15.D_Wp	real	0.000	13,613.630	52,366.400	19,193.104	22	9
X15.9.F.4_On.Stream	real	0.000	612.068	744.000	176.639	98	5
X15.9.F.4_qwi_boepd	real	0.000	32,607.951	54,637.613	11,030.283	103	1
X15.9.F.5_On.Stream	real	0.000	579.332	744.000	218.296	94	8
X15.9.F.5_qgp_boepd	real	0.000	0.210	0.304	0.126	6	1
X15.9.F.5_qo_boepd	real	0.000	1,406.055	2,026.048	846.673	6	1
X15.9.F.5_qwi_boepd	real	0.000	29,765.861	54,824.330	12,150.761	94	5
X15.9.F.5_qwp_boepd	real	0.000	463.017	801.447	258.079	6	1

Boolean Columns

name	data_type	min	mean	max	std	freq of max value
X15.9.F.1.C_qwi_boepd	bool	None	None	None	None	112
X15.9.F.1.C_Wi	bool	None	None	None	None	112

name	data_type	min	mean	max	std	freq of max value
X15.9.F.11_qwi_boepd	bool	None	None	None	None	112
X15.9.F.11_Wi	bool	None	None	None	None	112
X15.9.F.12_qwi_boepd	bool	None	None	None	None	112
X15.9.F.12_Wi	bool	None	None	None	None	112
X15.9.F.14_qwi_boepd	bool	None	None	None	None	112
X15.9.F.14_Wi	bool	None	None	None	None	112
X15.9.F.15.D_qwi_boepd	bool	None	None	None	None	112
X15.9.F.15.D_Wi	bool	None	None	None	None	112
X15.9.F.4_Gp	bool	None	None	None	None	112
X15.9.F.4_Np	bool	None	None	None	None	112
X15.9.F.4_qgp_boepd	bool	None	None	None	None	112
X15.9.F.4_qo_boepd	bool	None	None	None	None	112
X15.9.F.4_qwp_boepd	bool	None	None	None	None	112
X15.9.F.4_Wi	bool	None	None	None	None	112
X15.9.F.4_Wp	bool	None	None	None	None	112
X15.9.F.5_Gp	bool	None	None	None	None	112
X15.9.F.5_Np	bool	None	None	None	None	112
X15.9.F.5_Wi	bool	None	None	None	None	112
X15.9.F.5_Wp	bool	None	None	None	None	112

Shifts Detected

Driverless AI can perform shift detection between the training, validation and testing datasets. It does this by training a binomial model to predict which dataset a record belongs to. For example, it may find that it is able to separate the training and testing data with an AUC of 0.8 using only the column: `C1` as the predictor. This indicates that there is some sort of drift in the distribution of `C1` between the training and testing data.

For this experiment, Driverless AI was not able to check for distribution shifts because only the training dataset was supplied by the user.

Methodology

This section describes the experiment methodology.

Assumptions and Limitations

Driverless AI trains all models based on the training data provided (in this case: `make2p2i_train.csv.1543925877.14501.bin`). It is the assumption of Driverless AI that this dataset is representative of the data that will be seen when scoring.

Driverless AI may perform shift detection between the train data and another dataset. If a shift in distribution is detected, this may indicate that the data that will be used for scoring may have distributions not represented in the training data.

For this experiment, Driverless AI was not able to detect any shift in distribution between train data and another dataset because no validation or test data was provided.

Experiment Pipeline

For this experiment, Driverless AI performed the following steps to find the optimal final model:

The steps in this pipeline are described in more detail below:

1. **Ingest Data**
    - detected column types
2. **Feature Preprocessing**
    - turned raw features into numeric
3. **Model and Feature Tuning**
    - found the optimal parameters for xgboost and light gbm models by training models with different parameters
    - the best parameters are those that generate the least **RMSE** on the internal validation data
    - 83 of 86 models trained and scored to evaluate features and model parameters
4. **Feature Evolution**
    - found the best representation of the data for the final model training by creating and evaluating **1,499** features over **123** iterations
    - 160 of 168 models trained and scored to further evaluate engineered features
5. **Final Model**
    - the final model is the best model from the feature engineering iterations
    - no stacked ensemble is done because a time column was provided
6. **Create Scoring Pipeline**
    - created and exported the Python scoring pipeline (no MOJO Scoring Pipeline automatically created)
    - Python Scoring Pipeline: `h2oai_experiment_bosemoke/scoring_pipeline/scorer.zip`

Driverless AI trained models throughout the experiment in an effort to determine the best parameters, model dataset, and optimal final model. The stages are described below:

Driverless AI Stage	Timing	Number of Models
Data Preparation	2.84 secs	None
Model and Feature Tuning	525.93 secs	83 of 86 models
Feature Evolution	721.73 secs	160 of 168 models
Final Pipeline Training	4.33 secs	1 model

Experiment Settings

Below are the settings selected for the experiment by h2oai:

Defined Parameters

Parameter	Value
dataset_key	gicagowu
resumed_model_key	
target_col	X15.9.F.12_qo_boepd
weight_col	
fold_col	
orig_time_col	mth_yr
time_col	mth_yr
is_classification	False
cols_to_drop	[]
validset_key	
testset_key	
enable_gpus	True
seed	False
accuracy	10
time	4
interpretability	8
scorer	RMSE
time_groups_columns	['mth_yr']
num_prediction_periods	34
num_gap_periods	0
is_timeseries	True

Config Overrides

Parameter	Value
enable_xgboost	"auto"
enable_lightgbm	"auto"
enable_glm	"auto"
enable_tensorflow	"false"
enable_rulefit	"off"
check_distribution_shift	true
drop_features_distribution_shift_threshold_auc	0.6
enable_target_encoding	true
time_series_recipe	true

Parameter	Value
override_lag_sizes	""
prob_lag_non_targets	0.1
make_python_scoring_pipeline	true
make_mojo_scoring_pipeline	false
feature_brain_level	2
smart_imbalanced_sampling	true
holiday_features	true
seed	1234
max_orig_cols_selected	10000
nfeatures_max	-1
max_rows_feature_evolution	1000000
feature_engineering_effort	5
max_feature_interaction_depth	8
max_relative_cardinality	0.95
string_col_as_text_threshold	0.3
enable_tensorflow_force	false
tensorflow_max_epochs	100
enable_tensorflow_nlp	true
tensorflow_max_epochs_nlp	2
min_dai_iterations	0
max_nestimators	3000
max_learning_rate	0.05
max_cores	0
num_gpus_per_model	1
num_gpus_per_experiment	-1
gpu_id_start	0

These Accuracy, Time, and Interpretability settings map to the following internal configuration of the Driverless AI experiment:

Internal Parameter	Value
data filtered	False
tune target transform	True
number of feature engineering iterations	40
number of models trained per iteration	8
early stopping rounds	5

Internal Parameter	Value
monotonicity constraint	True
number of model tuning model combinations	81
number of base learners in ensemble	0
time column	mth_yr
time group columns	['mth_yr']
time period	2678400 seconds
number of prediction periods	34
number of gap periods	0

Details

- **data filtered**: Driverless AI may filter the training data depending on the number of rows and the Accuracy setting.
    - for this experiment, the training data was not filtered.
- **tune target transform**: whether Driverless AI evaluated the model performance if the target was transformed.
    - ex: the model performance may be better by predicting the log of the target column instead of the raw target column
- **number of feature engineering iterations**: the number of iterations performed of feature engineering.
- **number of models evaluated per iteration**: for each feature engineering iteration, Driverless AI trains multiple models. Each model is trained with a different set of predictors or features. The goal of this step is to determine which types of features, lead to the least RMSE.
- **early stopping rounds**: if Driverless AI does not see any improvement after 5 iterations of feature engineering, the feature engineering step is automatically stopped.
- **monotonicity constraint**: if enabled, the models will only have monotone relationships between the predictors and target variable.
- **number of model tuning combinations**: the number of model tuning combinations evaluated to determine the optimal model settings for the xgboost and light gbm models.
- **number of base learners in ensemble**: the number of base models used to create the final ensemble.
- **time column**: the column that provides time column. If a time column is provided, feature engineering and model validation will respect the causality of time. If the time column is turned off, no time order is used for modeling and data may be shuffled randomly (any potential temporal causality will be ignored).
- **time group columns**: the columns that make up the time series groups.
- **time period**: the periodicity found in the dataset.
- **number of prediction periods**: the number of periods you want to predict in advance.
- **number of gap periods**: the gap between the data available and the forecast period desired.

Validation Strategy

Driverless AI automatically split the data into training and validation data, ordering the data by `mth_yr`. The experiment predicted 34 **2678400 secondss** ahead with no gap between training and forecasting.

Model Tuning

The table below shows the score and training time of the xgboost and light gbm models evaluated by Driverless AI. The table shows the top 10 parameter tuning models evaluated, ordered based on a combination of least score and lowest training time.

job order	booster	nfeatures	scores	training times
53	gbtree	3	422.036	6.660
70	gbtree	3	433.676	10.830
79	gbtree	13	544.881	2.912
71	gbtree	15	548.104	4.601
62	gbtree	17	553.786	3.567
57	gbtree	3	556.783	1.057
0	lightgbm	5	569.040	1.967
72	gbtree	18	576.170	5.168
43	lightgbm	3	598.808	2.875
33	lightgbm	13	600.146	1.665

More detailed information on the parameters evaluated for each algorithm is shown below.

gbtree tuning

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
gpu_hist	lossguide	0.000	32.000	0.800	1.000	3	422.036	6.660
gpu_hist	lossguide	0.000	128.000	0.900	1.000	3	433.676	10.830
gpu_hist	lossguide	0.000	1,024.000	0.500	1.000	13	544.881	2.912
gpu_hist	depthwise	9.000	0.000	0.600	0.800	15	548.104	4.601
gpu_hist	lossguide	0.000	64.000	0.550	0.600	17	553.786	3.567
gpu_hist	lossguide	0.000	16.000	0.700	1.000	3	556.783	1.057
gpu_hist	lossguide	0.000	1,024.000	0.400	0.800	18	576.170	5.168
gpu_hist	lossguide	0.000	1,024.000	0.450	0.700	16	643.540	3.700
gpu_hist	lossguide	0.000	1,024.000	0.650	0.700	7	438.972	4.236
gpu_hist	depthwise	9.000	0.000	0.350	0.800	17	647.867	4.052

lightgbm tuning

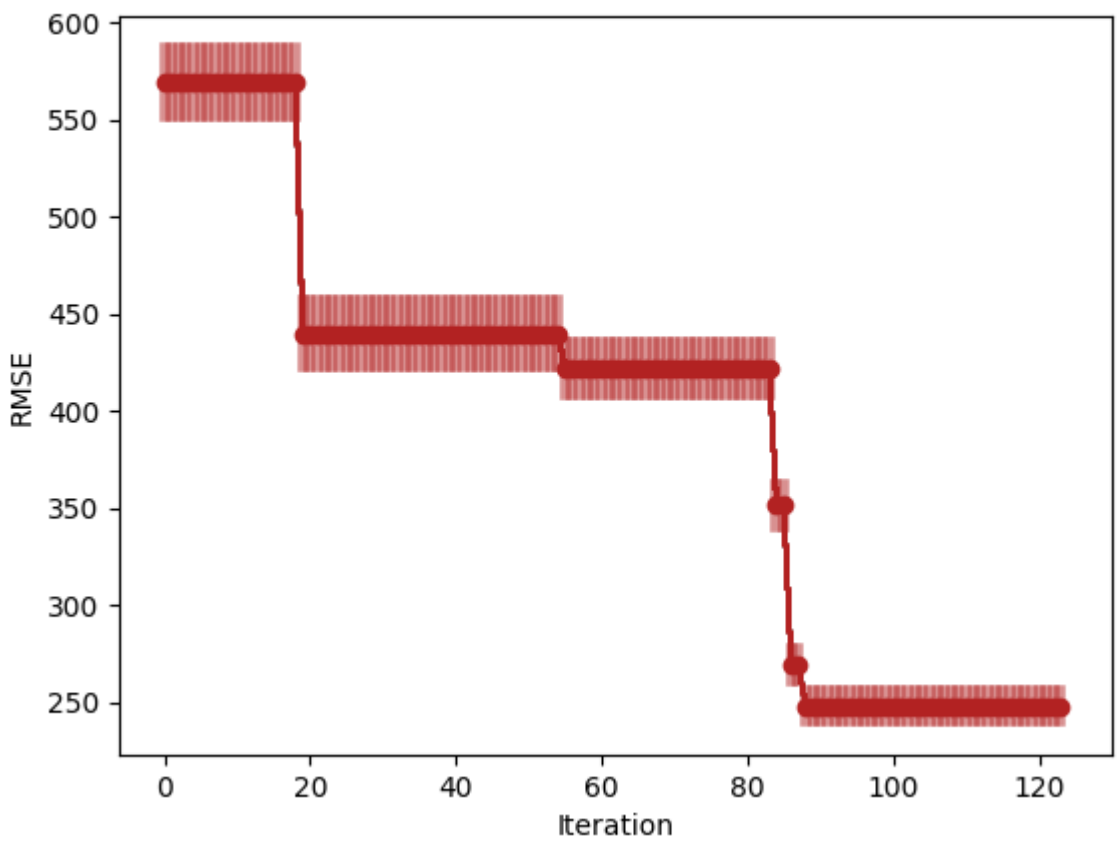
tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
gpu_hist	lossguide	0.000	64.000	0.800	0.700	5	569.040	1.967
gpu_hist	depthwise	5.000	0.000	0.900	0.500	3	598.808	2.875
gpu_hist	depthwise	10.000	0.000	0.500	0.800	13	600.146	1.665
gpu_hist	lossguide	0.000	4,096.000	0.700	0.900	8	602.752	2.295
gpu_hist	lossguide	0.000	64.000	0.300	1.000	19	665.425	2.241

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
gpu_hist	lossguide	0.000	16.000	0.700	0.700	8	671.754	2.132
gpu_hist	depthwise	8.000	0.000	0.800	1.000	6	692.626	2.622
gpu_hist	lossguide	0.000	64.000	0.600	0.700	18	731.764	4.754
gpu_hist	depthwise	8.000	0.000	0.650	0.600	5	756.599	1.932
gpu_hist	depthwise	7.000	0.000	0.700	0.900	4	439.813	1.727

Feature Evolution

During the Model and Feature Tuning Stage, Driverless AI evaluates the effects of different types of algorithms, algorithm parameters, and features. The goal of the Model and Feature Tuning Stage is to determine the best algorithm and parameters to use during the Feature Evolution Stage. In the Feature Evolution Stage, Driverless AI trained xgboost and light gbm models (160 of 168 models) where each model evaluated a different set of features. The Feature Evolution Stage uses a genetic algorithm to search the large feature engineering space.

The graph belows shows the effect the Model and Feature Tuning Stage and Feature Evolution Stage had on the performance.



Feature Transformation

The result of the Feature Evolution Stage is set of features to use for the final model. Some of these features were automatically created by Driverless AI. All 2 features used in the final model are shown belowed ordered by importance. If no transformer was applied, the feature is an original column.

Feature	Description	Transformer	Relative Importance
11_X15_9_F_12_qgp_boepd	X15_9_F_12_qgp_boepd (original)	None	1.000

Feature	Description	Transformer	Relative Importance
135_InteractionMul: X15_9_F_12_On_Stream: X15_9_F_12_qgp_boepd	[X15_9_F_12_On_Stream] * [X15_9_F_12_qgp_boepd]	Interaction	0.855

Final Model

Pipeline

Final XGBoostModel pipeline with ensemble_level=0 transforming 5 original features -> 2 features in each of 1 models each fit on time-based hold-out.

Details

Model Index	Type	Model Weight	Fitted features	Target Transformer
0	XGBoostModel	1	2	unit_box

- Model Index: 0 has a weight of 1 in the final ensemble

Type	learning rate	subsample	max_depth	tree_method	max_leaves	colsample_bytree	Split Type	grow_policy
XGBoostModel	0.020	1	9	gpu_hist	32	0.800	None	lossguide

Performance of Final Model

Scorer	Final (best individual) validation scores +/- standard deviation	Optimized	Better score is
RMSE	248.14 +/- 7.993	*	lower

Deployment

For this experiment, the Python Scoring Pipeline is available for productionizing the final model pipeline for a given row of data or table of data. The MOJO Scoring Pipeline can be built by clicking the **BUILD MOJO SCORING PIPELINE** button if available.

Python Scoring Pipeline

This package contains an exported model and Python 3.6 source code examples for productionizing models built using H2O Driverless AI. The Python Scoring Pipeline is located here:

- `h2oai_experiment_bosemoke/scoring_pipeline/scorer.zip`

The files in this package allow you to transform and score on new data in a couple of different ways:

- From Python 3.6, you can import a scoring module, and then use the module to transform and score on new data.
- From other languages and platforms, you can use the TCP/HTTP scoring service bundled with this package to call into the scoring pipeline module through remote procedure calls (RPC).