Driverless AI Experiment: `nucodewo`

Generated on: Sat Nov 24 22:36:06 2018

Generated by: h2oai

Table of Contents

Experiment Overview

Driverless AI built a stacked ensemble of 1 XGBoostModel, 1 GLMModel to predict `Drug_Overdose_Mortality_Rate` given 31 original features from the input dataset `100_Best_Features.csv`. This regression experiment completed in 1 hours and 33 minutes (1:33:13), using 27 of the 31 original features, and 4 of the 9,385 engineered features.

Performance

Dataset	RMSE
Internal Validation	5.02
Test Data	Test Data not Provided

Driverless Settings

Dial Settings	Description	Setting Value	Range of Possible Values
Accuracy	Controls sophistication of the model	7	1-10
Time	Controls duration of the experiment	6	1-10
Interpretability	Controls complexity of the features	6	1-10

System Specifications

System	System Memory	CPUs	GPUs
Linux	15	4	0

Versions

Driverless AI Version
1.3.1

Data Overview

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

data	file path	number of rows	number of columns
training	./tmp/hegiciko/100_Best_Features.csv	3,141	103
validation	Not provided		
testing	Not provided		

Training Data

The training data consists of both numeric and non numeric columns.

The summary of the columns is shown below:

Numeric Columns

name	min	mean	max	std	unique	freq of mode
Health_Factors_pctile_within_state	1.000	51.288	100.000	28.580	104	59
Length_of_Life_pctile_within_state	1.000	51.148	100.000	28.526	104	63
pct_Smokers	6.900	18.399	41.200	3.788	230	53
pct_Physically_Inactive	9.100	27.359	41.700	5.414	290	36
pct_Excessive_Drinking	8.400	16.574	27.300	3.346	176	45
Preventable_Hospitalization_rate	12.230	63.889	260.580	24.615	2,498	52
pct_Some_College	2.703	56.278	100.000	11.649	3,130	5
pct_Frequent_Mental_Distress	6.600	11.229	19.200	2.078	117	70
pct_Diabetic	5.100	11.137	22.800	2.282	148	73
Median_Household_Income	21,658.000	47,117.423	125,635.000	12,099.241	3,003	3
pct_Rural	0.000	58.624	100.000	31.505	2,413	701
NSSATS_COMPSAT_3	0.000	0.104	0.200	0.042	50	254
NMHSS_TREATFAMTHRPY	0.519	0.705	0.889	0.100	51	254
NMHSS_TREATTRAUMATHRPY	0.444	0.701	0.931	0.114	51	254
NMHSS_FOCUS_Mental health treatment	0.314	0.645	0.872	0.139	50	254

name	min	mean	max	std	unique	freq of mode
NMHSS_FOCUS_Mix of mental health and substance abuse treatment	0.111	0.323	0.686	0.138	51	254
NMHSS_LANGPROV_BOTH staff and on-call interpreter	0.011	0.213	0.566	0.135	51	254
NMHSS_HLTHREC_-6	0.000	0.038	0.104	0.025	49	274
NMHSS_FACNUM_11 to 30 facilities	0.000	0.001	0.009	0.002	12	2,304
NMHSS_IPSEXPERF_3	0.000	0.011	0.030	0.009	36	616
NMHSS_IPAGEPER017_1	0.000	0.005	0.016	0.005	27	1,156
NMHSS_IPETHPERHISP_2	0.000	0.011	0.049	0.011	35	802
NMHSS_IPETHPERNONHISP_4	0.000	0.001	0.024	0.003	8	2,650
NMHSS_IPRACETOTASIAN_41 to 50	0.000	0.000	0.003	0.001	3	2,825
NMHSS_IPRACEPERWHIT_1	0.000	0.000	0.006	0.001	5	2,699
NMHSS_IPRACEPERWHIT_6	0.000	0.026	0.089	0.017	41	274
NMHSS_IPRACEPERUNK_2	0.000	0.003	0.024	0.005	21	1,863
NMHSS_IPRACEPERUNK_3	0.000	0.002	0.013	0.003	17	1,913
NMHSS_OPAGETOT017_51 to 75	0.000	0.057	0.122	0.027	48	254
NMHSS_OPAGETOT017_None	0.090	0.212	0.533	0.063	51	254
NMHSS_OPAGEPER017_0	0.090	0.213	0.533	0.063	51	254
NMHSS_OPETHPERHISP_3	0.000	0.026	0.085	0.023	43	328
NMHSS_OPETHTOTUNK_1 to 10	0.000	0.086	0.271	0.053	50	254
NMHSS_OPRACEPERBLK_1	0.000	0.178	0.328	0.096	50	254
NMHSS_OPRACETOTWHIT_More than 250	0.000	0.072	0.146	0.035	50	254
NMHSS_OPRACEPERWHIT_7	0.000	0.215	0.608	0.117	51	254
NMHSS_OPLEGALTOTNONFOREN_11	0.000	0.001	0.006	0.002	7	2,574
NMHSS_PERCENTVA_7	0.000	0.000	0.005	0.001	5	2,834
NSDUH_PRUD_PYR_AVEG	0.006	0.008	0.010	0.001	51	254
NSDUH_SMI_PYR_AVEG	0.036	0.047	0.060	0.006	51	254
NSDUH_AMI_PYR_AVEG	0.168	0.195	0.242	0.017	51	254
TEDSA_AGE_8	0.085	0.121	0.146	0.011	52	254
TEDSA_DETNLF_3	0.000	0.085	0.361	0.054	51	254
TEDSA_ARRESTS_-9	0.000	0.054	0.975	0.148	51	254
TEDSA_ARRESTS_0	0.022	0.858	0.965	0.140	52	254

name	min	mean	max	std	unique	freq of mode
TEDSA_METHUSE_-9	0.000	0.047	1.000	0.161	41	514
TEDSA_SUB1_7	0.018	0.103	0.285	0.058	52	254
TEDSA_SUB2_7	0.016	0.054	0.136	0.025	52	254
TEDSA_FRSTUSE2_10	0.001	0.005	0.048	0.006	52	254
TEDSA_FRSTUSE2_11	0.000	0.003	0.043	0.005	52	254
TEDSA_FRSTUSE2_12	0.000	0.002	0.054	0.007	52	254
TEDSA_FRSTUSE2_7	0.009	0.023	0.110	0.013	52	254
TEDSA_FRSTUSE2_8	0.004	0.012	0.062	0.007	52	254
TEDSA_FRSTUSE2_9	0.002	0.008	0.061	0.007	52	254
TEDSA_SUB3_7	0.000	0.022	0.044	0.009	52	254
TEDSA_FRSTUSE3_-9	0.023	0.746	1.000	0.126	52	254
TEDSA_FRSTUSE3_10	0.000	0.003	0.063	0.008	52	254
TEDSA_FRSTUSE3_11	0.000	0.002	0.055	0.007	50	254
TEDSA_FRSTUSE3_12	0.000	0.002	0.067	0.009	48	254
TEDSA_FRSTUSE3_5	0.000	0.023	0.126	0.015	52	254
TEDSA_FRSTUSE3_6	0.000	0.018	0.135	0.017	52	254
TEDSA_FRSTUSE3_7	0.000	0.011	0.133	0.017	52	254
TEDSA_FRSTUSE3_8	0.000	0.006	0.096	0.012	52	254
TEDSA_FRSTUSE3_9	0.000	0.004	0.075	0.010	52	254
TEDSA_OPSYNFLG_0	0.578	0.826	0.957	0.081	52	254
TEDSA_OPSYNFLG_1	0.043	0.174	0.422	0.081	52	254
TEDSA_OTCFLG_0	0.990	0.997	1.000	0.002	51	254
TEDSA_OTCFLG_1	0.000	0.003	0.010	0.002	51	254
TEDSA_DSMCRIT_-9	0.003	0.598	1.000	0.332	41	802
TEDSA_DSMCRIT_12	0.000	0.004	0.025	0.006	39	841
TEDSA_DSMCRIT_14	0.000	0.004	0.053	0.009	24	1,782
TEDSA_DSMCRIT_15	0.000	0.009	0.116	0.019	25	1,680
TEDSA_DSMCRIT_17	0.000	0.004	0.054	0.010	25	1,680
TEDSA_DSMCRIT_18	0.000	0.001	0.012	0.002	24	1,782
TEDSA_DSMCRIT_2	0.000	0.017	0.199	0.042	36	1,065
TEDSA_DSMCRIT_5	0.000	0.073	0.414	0.077	39	841
TEDSA_PSYPROB_1	0.000	0.269	0.695	0.187	48	561
TEDSD_AGE_8	0.083	0.112	0.148	0.011	51	271

name	min	mean	max	std	unique	freq of mode
TEDSD_DETNLF_1	0.000	0.011	0.058	0.011	50	271
TEDSD_PREG_1	0.003	0.015	0.041	0.007	51	271
TEDSD_CBSA_14540	0.000	0.002	0.061	0.012	8	2,520
TEDSD_CBSA_17300	0.000	0.006	0.145	0.028	8	2,520
TEDSD_CBSA_21060	0.000	0.001	0.020	0.004	8	2,520
TEDSD_CBSA_21780	0.000	0.000	0.003	0.001	8	2,520
TEDSD_CBSA_26580	0.000	0.004	0.091	0.017	8	2,520
TEDSD_CBSA_30460	0.000	0.003	0.073	0.014	8	2,520
TEDSD_CBSA_30940	0.000	0.000	0.002	0.000	8	2,520
TEDSD_CBSA_31140	0.000	0.005	0.119	0.023	8	2,520
TEDSD_CBSA_36980	0.000	0.001	0.022	0.004	8	2,520
TEDSD_DIVISION_6	0.000	0.103	1.000	0.286	8	2,358
TEDSD_SUB2_7	0.009	0.053	0.134	0.026	51	271
TEDSD_SUB3_7	0.000	0.022	0.051	0.010	51	271
TEDSD_DSMCRIT_14	0.000	0.005	0.060	0.013	27	1,555
TEDSD_PSYPROB_1	0.000	0.261	0.687	0.178	45	614
RetailDrug_AMOBARBITAL (SCHEDULE 2)	0.460	27.118	152.210	28.700	41	295
RetailDrug_AMPHETAMINE	25,380.020	544,943.799	1,534,182.940	410,506.389	51	254
RetailDrug_FENTANYL BASE	397.690	10,437.374	34,415.750	8,312.239	51	254
RetailDrug_LISDEXAMFETAMINE	5,188.460	290,112.458	1,086,524.420	270,318.638	51	254
RetailDrug_METHYLPHENIDATE (DL;D;L;ISOMERS)	32,818.620	503,090.213	1,353,418.140	367,019.951	51	254
RetailDrug_TAPENTADOL	5,498.000	169,584.071	480,772.940	146,550.897	51	254
Drug_Overdose_Mortality_Rate	2.713	17.287	84.882	6.646	1,587	717

Non-Numeric Columns

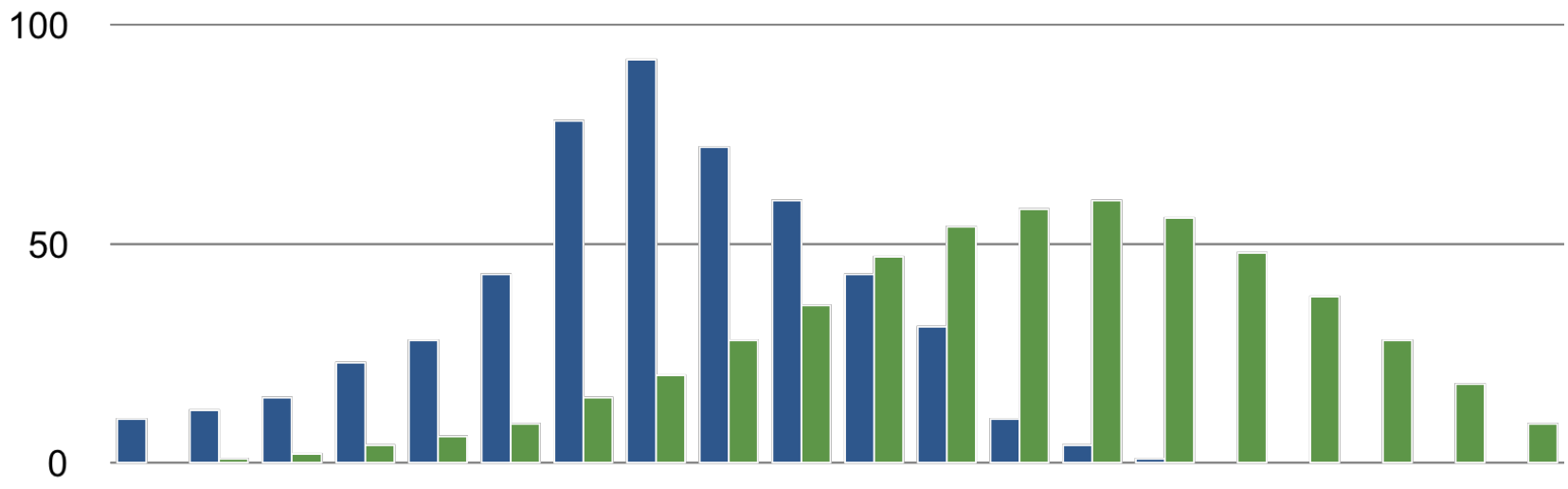
name	unique	top	freq of top value
STATE	51	TX	254
County	1,848	Washington	31

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.

An example of a shift distribution between two datasets is shown below:



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

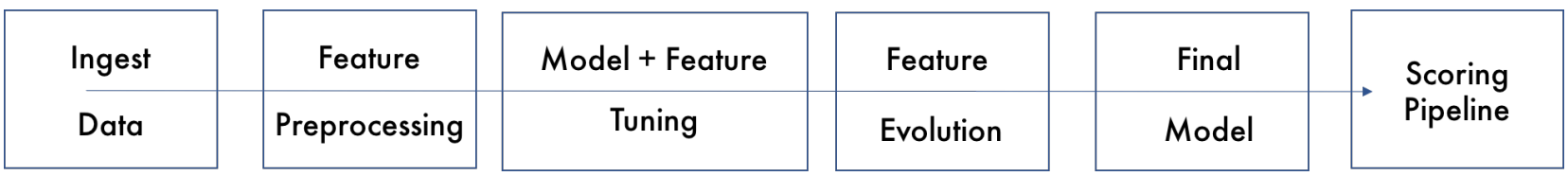
Driverless AI trains all models based on the training data provided (in this case: `100_Best_Features.csv`). It is the assumption of Driverless AI that this dataset is representative of the data that will be seen when scoring.

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 models by training models with different parameters
    - the best parameters are those that generate the least **RMSE** on the internal validation data
    - trained and scored **41** models 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 **9,385** features over **100** iteration s
    - trained and scored **808** models to further evaluate engineered features
5. **Final Model**
    - the final model is a stacked ensemble of **1 XGBoostModel, 1 GLMModel**.
    - the features of these models are the best features found during the feature engineering iterations
6. **Create Scoring Pipeline**
    - created and exported the Python scoring pipeline (no MOJO Scoring Pipeline created)
    - Python Scoring Pipeline: `h2oai_experiment_nucodewo/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 Step	Number of Models	Number of Folds/Validation Datasets
Parameter and Feature Tuning	41	3
Feature Evolution	808	3
Final Model	2	3

Experiment Settings

Below are the settings selected for the experiment by h2oai:

Defined Parameters

Parameter	Value
dataset_key	hegiciko
target_col	Drug_Overdose_Mortality_Rate
weight_col	
fold_col	
orig_time_col	
time_col	[OFF]
is_classification	False
cols_to_drop	[]
validset_key	
testset_key	
enable_gpus	False

Parameter	Value
seed	False
accuracy	7
time	6
interpretability	6
scorer	RMSE
is_timeseries	False

Config Overrides

Parameter	Value
enable_xgboost	"auto"
enable_glm	"auto"
enable_tensorflow	"off"
enable_rulefit	"off"
enable_lightgbm	"off"
check_distribution_shift	true
time_series_recipe	true
make_python_scoring_pipeline	true
make_mojo_scoring_pipeline	false
smart_imbalanced_sampling	false
seed	1234
nfeatures_max	-1
feature_engineering_effort	5
max_feature_interaction_depth	8
max_relative_cardinality	0.95
string_col_as_text_threshold	0.3
tensorflow_max_epochs	100
tensorflow_max_epochs_nlp	2
max_nestimators	3000
max_learning_rate	0.5
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	100
number of models trained per iteration	4
early stopping rounds	10
monotonicity constraint	False
number of model tuning model combinations	41
number of base learners in ensemble	2
time column	[OFF]

Details

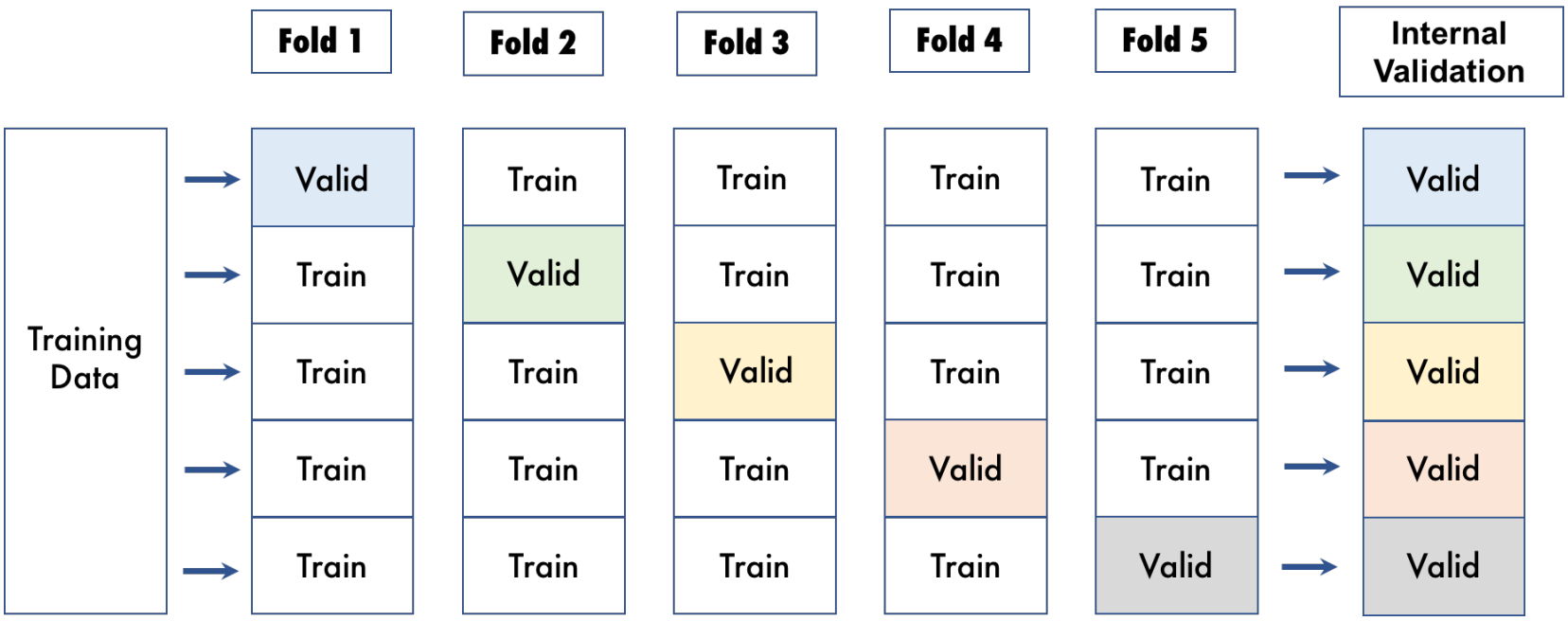
- **data filtered**: Driverless AI may filter the training data depending on the number of rows and the Accuracy setting.
    - for this experiment, the training data was not filtered.
- **tune target transform**: whether Driverless AI evaluated the model performance if the target was transformed.
    - 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 10 iterations of feature engineering, the feature engineering step is automatically stopped.
- **monotonicity constraint**: if enabled, the xgboost 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 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).

Validation Strategy

Driverless AI automatically split the training data to determine the performance of the model parameter tuning and feature engineering steps.

For the experiment, Driverless AI randomly split the data into 3 fold cross validation. With cross validation, the whole dataset is utilized by training 3 models where each model is trained on a different subset of the training data.

The visualization below shows how cross validation is utilized to get predictions on hold out data. The visualization shows an example of cross validation with 5 folds. For this experiment, however, 3 folds were created.



Model Tuning

The table below shows a portion of the different parameter configurations evaluated by Driverless AI for the xgboost models and their score and training time. The table is ordered based on a combination of least score and lowest training time.

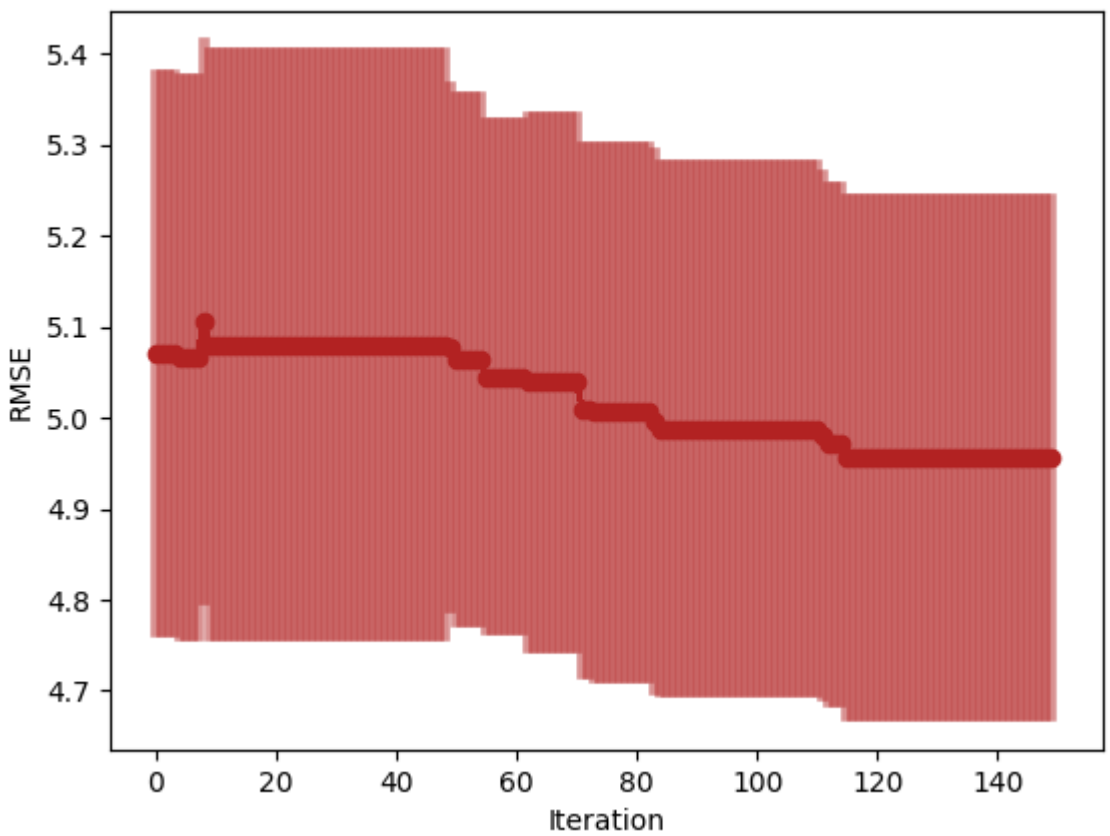
tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
hist	depthwise	6.000	0.000	0.900	0.500	95	5.080	8.684
hist	depthwise	6.000	0.000	0.900	0.500	95	5.080	8.497
hist	depthwise	4.000	0.000	0.650	0.700	95	5.135	4.798
hist	lossguide	0.000	16.000	0.800	0.900	254	5.139	12.743
hist	depthwise	4.000	0.000	0.650	0.700	268	5.145	13.748
hist	depthwise	3.000	0.000	0.800	0.600	91	5.148	4.020
hist	lossguide	0.000	512.000	0.650	0.800	280	5.150	373.911
hist	depthwise	10.000	0.000	0.300	0.900	99	5.157	31.469
hist	depthwise	10.000	0.000	0.900	0.500	95	5.157	20.219
hist	depthwise	10.000	0.000	0.900	0.500	95	5.157	20.358
hist	depthwise	10.000	0.000	0.900	0.500	86	5.160	18.825
hist	depthwise	10.000	0.000	0.900	0.500	86	5.160	18.788
hist	depthwise	10.000	0.000	0.400	0.800	300	5.092	97.753
hist	lossguide	0.000	64.000	0.800	1.000	91	5.166	11.362
hist	lossguide	0.000	16.000	0.300	0.700	308	5.167	11.392
hist	lossguide	0.000	256.000	0.450	0.400	317	5.168	68.692
hist	depthwise	8.000	0.000	0.550	0.700	97	5.171	16.814
hist	depthwise	6.000	0.000	0.600	0.900	295	5.172	23.500
hist	depthwise	9.000	0.000	0.350	0.700	301	5.178	62.244

tree method	grow policy	max depth	max leaves	colsample bytree	subsample	nfeatures	scores	training times
hist	depthwise	9.000	0.000	0.500	0.900	286	5.181	87.849
hist	lossguide	0.000	512.000	0.500	1.000	282	5.183	111.029
hist	lossguide	0.000	64.000	0.300	0.500	321	5.187	30.391
hist	depthwise	5.000	0.000	0.200	0.500	307	5.197	10.369
hist	depthwise	6.000	0.000	0.900	0.500	90	5.097	8.902
hist	depthwise	7.000	0.000	0.300	0.800	310	5.200	23.817
hist	lossguide	0.000	32.000	0.350	0.900	100	5.202	5.823
hist	lossguide	0.000	1,024.000	0.450	0.700	291	5.207	204.957
hist	lossguide	0.000	64.000	0.800	0.800	277	5.215	44.489
hist	depthwise	10.000	0.000	0.900	0.500	280	5.221	83.225
hist	lossguide	0.000	1,024.000	0.200	0.700	293	5.277	29.028
hist	lossguide	0.000	1,024.000	0.800	1.000	258	5.279	304.015
hist	lossguide	0.000	64.000	0.650	0.500	307	5.293	18.230
hist	depthwise	10.000	0.000	0.700	1.000	267	5.308	127.064
hist	lossguide	0.000	16.000	0.200	0.600	101	5.339	4.198
hist	depthwise	6.000	0.000	0.900	0.500	90	5.097	9.337
hist	lossguide	0.000	512.000	0.900	0.900	68	5.350	36.222
hist	depthwise	8.000	0.000	0.350	0.800	304	5.098	87.700
hist	lossguide	0.000	64.000	0.800	0.700	78	5.106	13.284
hist	depthwise	5.000	0.000	0.650	0.900	96	5.119	9.930
hist	depthwise	3.000	0.000	0.350	0.700	260	5.124	13.747
hist	depthwise	5.000	0.000	0.550	1.000	264	5.128	31.363

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 algorithm parameters to use during the Feature Evolution Stage. The Feature Evolution Stage trained 808 xgboost 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 the final set of features to use for the model. Some of these features were automatically created by Driverless AI. The top 14 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
3_Length_of_Life_pctile_within_state	Length_of_Life_pctile_within_state (original)	None	1.000
99_pct_Rural	pct_Rural (original)	None	0.562
53_TEDSA_DSMCRIT_5	TEDSA_DSMCRIT_5 (original)	None	0.170
25_NMHSS_OPRACEPERBLK_1	NMHSS_OPRACEPERBLK_1 (original)	None	0.159
4_Median_Household_Income	Median_Household_Income (original)	None	0.158
97_pct_Frequent_Mental_Distress	pct_Frequent_Mental_Distress (original)	None	0.152
100_pct_Smokers	pct_Smokers (original)	None	0.136
0_Freq: County	Encoding of categorical levels of feature(s) ['County'] to value between 0 and 1 based on their relative frequency	Frequency Encoding	0.104
64_TEDSA_FRSTUSE3_5	TEDSA_FRSTUSE3_5 (original)	None	0.093
29_NMHSS_TREATFAMTHRPY	NMHSS_TREATFAMTHRPY (original)	None	0.090
42_TEDSA_AGE_8	TEDSA_AGE_8 (original)	None	0.089
43_TEDSA_ARRESTS_-9	TEDSA_ARRESTS_-9 (original)	None	0.088

Feature	Description	Transformer	Relative Importance
96_pct_Excessive_Drinking	pct_Excessive_Drinking (original)	None	0.272
12_NMHSS_IPRACEPERUNK_2	NMHSS_IPRACEPERUNK_2 (original)	None	0.079

Final Model

Pipeline

Final StackedEnsemble pipeline with ensemble_level=2 transforming 29 original features -> 31 features in each of 2 models each fit on 3 internal holdout splits then linearly blended

Final Model Scores

Scorer	Final ensemble external validation scores +/- standard deviation	Optimized	Better score is
GINI	0.68861 +/- 0.02215		higher
R2	0.43681 +/- 0.062708		higher
MSE	25.313 +/- 2.9173		lower
RMSLE	0.24628 +/- 0.0078511		lower
RMSPE	0.31386 +/- 0.019966		lower
MAE	3.0992 +/- 0.11826		lower
MER	11.387 +/- 0.62292		lower
MAPE	19.579 +/- 0.82387		lower
SMAPE	17.936 +/- 0.60953		lower
RMSE	5.0203 +/- 0.29053	*	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_nucodewo/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).