# Driverless AI Experiment: Diabetes DataSet

Generated by: h2oai

Generated on: Fri Dec 13 03:38:06 2019

[Driverless AI Experiment: Diabetes DataSet 1](#_Toc20130252)

[Experiment Overview 1](#_Toc20130253)

[Data Overview 3](#_Toc20130254)

[Methodology 5](#_Toc20130255)

[Data Sampling 10](#_Toc20130256)

[Validation Strategy 10](#_Toc20130257)

[Model Tuning 11](#_Toc20130258)

[Feature Evolution 13](#_Toc20130259)

[Feature Transformations 13](#_Toc20130260)

[Final Model 15](#_Toc20130261)

[Alternative Models 21](#_Toc20130262)

[Deployment 22](#_Toc20130263)

[Partial Dependence Plots 23](#_Toc20130264)

[Appendix 25](#_Toc20130266)

## Experiment Overview

Driverless AI built a stacked ensemble of 1 XGBoostGBMModel to predict *readmitted* given 46 original features from the input dataset *dataset\_diabetes.zip*. This classification experiment completed in 17 minutes and 19 seconds (0:17:19), using 5 of the 46 original features, and 337 of the 1,699 engineered features.

### Performance

|  |  |
| --- | --- |
| **Dataset** | **AUC** |
| Internal Validation | 0.793 |
| Test Data | Test Data not Provided |

### Driverless Settings

|  |  |  |  |
| --- | --- | --- | --- |
| Dial Settings | Description | Setting Value | Range of Possible Values |
| Accuracy | Controls accuracy needs of the model | 5 | 1-10 |
| Time | Controls duration of the experiment | 5 | 1-10 |
| Interpretability | Controls complexity of the model | 5 | 1-10 |

### System Specifications

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Address** | **System** | **System Memory** | **CPUs** | **GPUs** |
| http://127.0.0.1:12345 | Docker/Linux | 240 | 32 | 4 |

### Versions

|  |
| --- |
| Driverless AI Version |
| 1.8.1 |

## Data Overview

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

|  |  |  |  |
| --- | --- | --- | --- |
| **data** | **file path** | **number of rows** | **number of columns** |
| training | ./tmp/9899bbb2-1d56-11ea-b321-0242ac110002/dataset\_diabetes.zip.1576206836.1184256.bin | 101,766 | 50 |
| validation | Not provided | None | None |
| testing | Not provided | None | None |

### Training Data

The training data consists of both numeric and categorical columns.

The summary of the columns is shown below:

#### Numeric Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **data\_type** | **min** | **mean** | **max** | **std** | **unique** | **freq of mode** |
| encounter\_id | int | 12,522.000 | 165,201,645.623 | 443,867,222.000 | 102,640,295.983 | 101,766 | 1 |
| patient\_nbr | int | 135.000 | 54,330,400.695 | 189,502,619.000 | 38,696,359.347 | 71,518 | 40 |
| admission\_type\_id | int | 1.000 | 2.024 | 8.000 | 1.445 | 8 | 53,990 |
| discharge\_disposition\_id | int | 1.000 | 3.716 | 28.000 | 5.280 | 26 | 60,234 |
| admission\_source\_id | int | 1.000 | 5.754 | 25.000 | 4.064 | 17 | 57,494 |
| time\_in\_hospital | int | 1.000 | 4.396 | 14.000 | 2.985 | 14 | 17,756 |
| num\_lab\_procedures | int | 1.000 | 43.096 | 132.000 | 19.674 | 118 | 3,208 |
| num\_procedures | int | 0.000 | 1.340 | 6.000 | 1.706 | 7 | 46,652 |
| num\_medications | int | 1.000 | 16.022 | 81.000 | 8.128 | 75 | 6,086 |
| number\_outpatient | int | 0.000 | 0.369 | 42.000 | 1.267 | 39 | 85,027 |
| number\_emergency | int | 0.000 | 0.198 | 76.000 | 0.930 | 33 | 90,383 |
| number\_inpatient | int | 0.000 | 0.636 | 21.000 | 1.263 | 21 | 67,630 |
| number\_diagnoses | int | 1.000 | 7.423 | 16.000 | 1.934 | 16 | 49,474 |

#### Categorical Columns

|  |  |  |  |
| --- | --- | --- | --- |
| **name** | **unique** | **top** | **freq of top value** |
| race | 5 | Caucasian | 76,099 |
| gender | 3 | Female | 54,708 |
| age | 10 | [70-80) | 26,068 |
| weight | 9 | [75-100) | 1,336 |
| payer\_code | 17 | MC | 32,439 |
| medical\_specialty | 72 | InternalMedicine | 14,635 |
| diag\_1 | 716 | 428 | 6,862 |
| diag\_2 | 748 | 276 | 6,752 |
| diag\_3 | 789 | 250 | 11,555 |
| max\_glu\_serum | 3 | Norm | 2,597 |
| A1Cresult | 3 | >8 | 8,216 |
| metformin | 4 | No | 81,778 |
| repaglinide | 4 | No | 100,227 |
| nateglinide | 4 | No | 101,063 |
| chlorpropamide | 4 | No | 101,680 |
| glimepiride | 4 | No | 96,575 |
| acetohexamide | 2 | No | 1 |
| glipizide | 4 | No | 89,080 |
| glyburide | 4 | No | 91,116 |
| tolbutamide | 2 | No | 23 |
| pioglitazone | 4 | No | 94,438 |
| rosiglitazone | 4 | No | 95,401 |
| acarbose | 4 | No | 101,458 |
| miglitol | 4 | No | 101,728 |
| troglitazone | 2 | No | 3 |
| tolazamide | 3 | No | 101,727 |
| examide | 1 | No | 101,766 |
| citoglipton | 1 | No | 101,766 |
| insulin | 4 | No | 47,383 |
| glyburide-metformin | 4 | No | 101,060 |
| glipizide-metformin | 2 | No | 13 |
| glimepiride-pioglitazone | 2 | No | 1 |
| metformin-rosiglitazone | 2 | No | 2 |
| metformin-pioglitazone | 2 | No | 1 |
| change | 2 | No | 54,755 |
| diabetesMed | 2 | Yes | 78,363 |
| readmitted | 3 | NO | 54,864 |

### 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: *dataset\_diabetes.zip*). 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:

* **Ingest Data** 
  + - detected column types
* **Feature Preprocessing**
  + turned raw features into numeric
* **Model and Feature Tuning**

This stage combines random hyperparameter tuning with feature selection and generation. Features in each iteration are updated using variable importance from the previous iteration as a probabilistic prior to decide what new features to create. The best performing model and features are then passed to the feature evolution stage.

* + found the optimal parameters for xgboost, linear and light gbm models by training models with different parameters
  + the best parameters are those that generate the greatest **AUC** on the internal validation data
  + 14 models trained and scored to evaluate features and model parameters
* **Feature Evolution**

This stage uses a genetic algorithm to find the best set of model parameters and feature transformations to be used in the final model.

* + found the best representation of the data for the final model training by creating and evaluating **1,699** features over **20** iterations
  + trained and scored 59 models to further evaluate engineered features
* **Final Model** 
  + - the final model is a stacked ensemble of **1 XGBoostGBMModel**
    - the features of this model are the best features found during the feature engineering iterations
* **Create Scoring Pipeline** 
  + created and exported the Python scoring pipeline (no MOJO Scoring Pipeline automatically created)
  + Python Scoring Pipeline: h2oai\_experiment\_825d312a-1d57-11ea-b321-0242ac110002/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 (seconds) | Number of Models |
| Data Preparation | 14.41 | 0 |
| Model and Feature Tuning | 146.98 | 14 |
| Feature Evolution | 622.69 | 59 |
| Final Pipeline Training | 141.17 | 4 |

### Experiment Settings

Below are the settings selected for the experiment by h2oai. The Defined Parameters represent the high-level parameters.

**Defined Parameters**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| is\_classification | True |
| enable\_gpus | True |
| seed | False |
| accuracy | 5 |
| time | 5 |
| interpretability | 5 |
| time\_groups\_columns | None |
| num\_prediction\_periods | None |
| num\_gap\_periods | None |
| is\_timeseries | False |

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

|  |  |
| --- | --- |
| **Internal Parameter** | **Value** |
| data filtered | False |
| number of feature engineering iterations | 50 |
| number of models trained per iteration | 8 |
| early stopping rounds | 10 |
| monotonicity constraint | False |
| number of model tuning model combinations | 13 |
| number of base learners in ensemble | 1 |
| 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.
* **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 greatest AUC.
* **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 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 the 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).

## Data Sampling

Driverless AI did not perform any down sampling of the data.

## 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 **2/3 training** and **1/3 validation**.

## Model Tuning

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **job order** | **booster** | **nfeatures** | **scores** | **training times** |
| 12 | lightgbm | 471 | 0.7905697215 | 11.3586313725 |
| 13 | gbtree | 373 | 0.7905048212 | 4.4413166046 |
| 7 | lightgbm | 114 | 0.7825442708 | 16.559073925 |
| 11 | lightgbm | 260 | 0.7802134815 | 15.2470033169 |
| 5 | gblinear | 44 | 0.7637110589 | 3.1049726009 |
| 1 | lightgbm | 114 | 0.7901216884 | 7.9689700603 |
| 3 | lightgbm | 37 | 0.7892815233 | 6.9257040024 |
| 0 | lightgbm | 114 | 0.7864768482 | 11.8405835629 |
| 2 | lightgbm | 36 | 0.7864594496 | 10.4607760906 |
| 9 | lightgbm | 114 | 0.7858413282 | 6.7498903275 |

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 | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 373 | 0.7905048212 | 4.4413166046 |
| gpu\_hist | depthwise | 9.0 | 0.0 | 0.35 | 0.4 | 114 | 0.7847939955 | 5.594363451 |
| gpu\_hist | lossguide | 0.0 | 512.0 | 0.2 | 0.7 | 316 | 0.7838915636 | 8.5557088852 |

### gblinear tuning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **reg alpha** | **reg lambda** | **n lambda** | **min lambda fraction** | **nfeatures** | **scores** | **training times** |
| 0.0005 | 0.001 | 1 | 0.0001 | 44 | 0.7637110589 | 3.1049726009 |

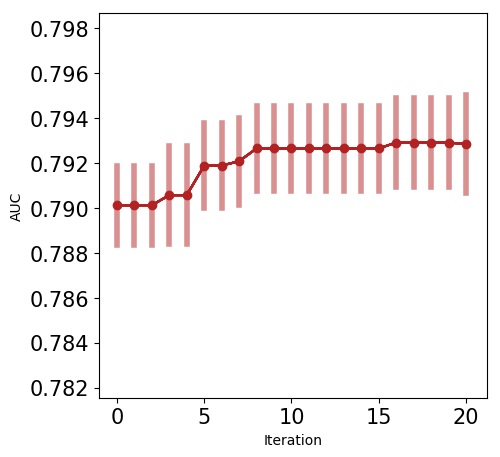
### lightgbm tuning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **tree method** | **grow policy** | **max depth** | **max leaves** | **colsample bytree** | **subsample** | **nfeatures** | **scores** | **training times** |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 471 | 0.7905697215 | 11.3586313725 |
| gpu\_hist | lossguide | 0.0 | 512.0 | 0.35 | 0.6 | 114 | 0.7825442708 | 16.559073925 |
| gpu\_hist | lossguide | 0.0 | 512.0 | 0.45 | 0.6 | 260 | 0.7802134815 | 15.2470033169 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 114 | 0.7901216884 | 7.9689700603 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 37 | 0.7892815233 | 6.9257040024 |
| gpu\_hist | depthwise | 10.0 | 0.0 | 0.8 | 0.7 | 114 | 0.7864768482 | 11.8405835629 |
| gpu\_hist | depthwise | 10.0 | 0.0 | 0.8 | 0.7 | 36 | 0.7864594496 | 10.4607760906 |
| gpu\_hist | depthwise | 9.0 | 0.0 | 0.45 | 0.4 | 114 | 0.7858413282 | 6.7498903275 |
| gpu\_hist | lossguide | 0.0 | 512.0 | 0.8 | 0.7 | 114 | 0.7854737524 | 10.908217907 |

## 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 (59) 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 below shows the effect the Model and Feature Tuning Stage and Feature Evolution Stage had on the performance.



Based on the experiment settings and column types in the dataset, Driverless AI was able to explore the following transformers:

* **ClusterDistTransformer**: clusters selected numeric columns and uses the distance to a specific cluster as a new feature.
* **ClusterIdTransformer**: clusters selected numeric columns and uses the centroid assignment as a new feature.
* **ClusterTETransformer**: clusters selected numeric columns and calculates the mean of the response column for each cluster. The mean of the response is used as a new feature. Cross Validation is used to calculate mean response to prevent overfitting.
* **InteractionsTransformer**: add, divide, multiply, and subtract two numeric columns in the data to create a new feature.
* **NumToCatTETransformer**: converts numeric columns to categoricals by binning and then calculates the mean of the response column for each group. The mean of the response for the bin is used as a new feature. Cross Validation is used to calculate mean response to prevent overfitting.
* **NumToCatWoETransformer**: converts a numeric column to categorical by binning and then calculates Weight of Evidence for each bin. The Weight of Evidence is used as a new feature. Weight of Evidence measures the “strength” of a grouping for separating good and bad risk and is calculated by taking the log of the ratio of distributions for a binary response column.
* **TruncSVDNumTransformer**: trains a Truncated SVD model on selected numeric columns and uses the components of the truncated SVD matrix as new features.
* **CVTargetEncodeTransformer**: calculates the mean of the response column for each value in a categorical column and uses this as a new feature. Cross Validation is used to calculate mean response to prevent overfitting.
* **FrequentTransformer**: calculates the frequency for each value in categorical column(s) and uses this as a new feature.
* **WeightOfEvidenceTransformer**: calculates Weight of Evidence for each value in categorical column(s). The Weight of Evidence is used as a new feature. Weight of Evidence measures the “strength” of a grouping for separating good and bad risk and is calculated by taking the log of the ratio of distributions for a binary response column.
* **CatTransformer**: None
* **OneHotEncodingTransformer**: converts a categorical column to a series of boolean features by performing one-hot encoding. The boolean features are used as new features.
* **DatesTransformer**: retrieves any date or time values, including: Year, Quarter, Month, Day, Day of Year, Week, Weekday, Hour, Minute, Second.
* **IsHolidayTransformer**: determines if a date column is a holiday. A boolean column indicating if the date is a holiday is added as a new feature.
* **TextBiGRUTransformer**: trains a bi-directional GRU TensorFlow model on word embeddings created from a text feature to predict the response column. The GRU prediction is used as a new a feature. Cross Validation is used when training the GRU model to prevent overfitting.
* **TextCNNTransformer**: trains a CNN TensorFlow model on word embeddings created from a text feature to predict the response column. The CNN prediction is used as a new a feature. Cross Validation is used when training the CNN model to prevent overfitting.
* **TextCharCNNTransformer**: trains a CNN TensorFlow model on character embeddings created from a text feature to predict the response column. The CNN prediction is used as a new a feature. Cross Validation is used when training the CNN model to prevent overfitting.
* **TextLinModelTransformer**: trains a linear model on a TF-IDF matrix created from a text feature to predict the response column. The linear model prediction is used as a new feature. Cross Validation is used when training the linear model to prevent overfitting.
* **TextTransformer**: tokenizes a text column and creates a TFIDF matrix (term frequency-inverse document frequency) or count (count of the word) matrix. This may be followed by dimensionality reduction using truncated SVD. Selected components of the TF-IDF/Count matrix are used as new features.
* **CVCatNumEncodeTransformer**: calculates an aggregation of a numeric column for each value in a categorical column (ex: calculate the mean Temperature for each City) and uses this aggregation as a new feature.
* **NumCatTETransformer**: calculates the mean of the response column for several selected columns. If one of the selected columns is numeric, it is first converted to categorical by binning. The mean of the response column is used as a new feature. Cross Validation is used to calculate mean response to prevent overfitting.

## Feature Transformations

The result of the Feature Evolution Stage is a set of features to use for the final model. Some of these features were automatically created by Driverless AI. The top features used in the final model are shown below, ordered by importance. The features in the table are limited to the top 50, restricted to those with relative importance greater than or equal to 0.003. If no transformer was applied, the feature is an original column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Feature** | **Description** | **Transformer** | **Relative Importance** |
| 1 | 103\_ClusterTE: ClusterID10: discharge\_disposition\_id.0 | Out-of-fold mean of the response (class 0) grouped by: ['ClusterID10:discharge\_disposition\_id'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 1.0 |
| 2 | 47\_ClusterDist20: discharge\_disposition\_id: num\_medications: number\_diagnoses: number\_emergency: number\_inpatient: patient\_nbr.9 | Distances to cluster center after segmenting columns ['discharge\_disposition\_id', 'num\_medications', 'number\_diagnoses', 'number\_emergency', 'number\_inpatient', 'patient\_nbr'] into 20 clusters. Distance to cluster #9 | Cluster Distance | 0.7757 |
| 3 | 122\_NumToCatTE: number\_inpatient: patient\_nbr.2 | Out-of-fold mean of the response (class 2) grouped by: ['number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(100, 1, None)] | Numeric to Categorical Target Encoding | 0.7388 |
| 4 | 85\_NumToCatTE: patient\_nbr.2 | Out-of-fold mean of the response (class 2) grouped by: ['patient\_nbr'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 3, None)] | Numeric to Categorical Target Encoding | 0.7127 |
| 5 | 122\_NumToCatTE: number\_inpatient: patient\_nbr.1 | Out-of-fold mean of the response (class 1) grouped by: ['number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(100, 1, None)] | Numeric to Categorical Target Encoding | 0.6839 |
| 6 | 103\_ClusterTE: ClusterID10: discharge\_disposition\_id.1 | Out-of-fold mean of the response (class 1) grouped by: ['ClusterID10:discharge\_disposition\_id'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.5759 |
| 7 | 85\_NumToCatTE: patient\_nbr.1 | Out-of-fold mean of the response (class 1) grouped by: ['patient\_nbr'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 3, None)] | Numeric to Categorical Target Encoding | 0.5522 |
| 8 | 29\_CVTE: diag\_1.2 | Out-of-fold mean of the response (class 2) grouped by: ['diag\_1'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.5371 |
| 9 | 142\_NumToCatTE: admission\_source\_id: discharge\_disposition\_id: number\_inpatient: patient\_nbr.2 | Out-of-fold mean of the response (class 2) grouped by: ['admission\_source\_id', 'discharge\_disposition\_id', 'number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 10 equally populated bins) [internal parameters:(20, 3, 10)] | Numeric to Categorical Target Encoding | 0.5228 |
| 10 | 8\_number\_inpatient | number\_inpatient (Orig) | None | 0.5035 |
| 11 | 29\_CVTE: diag\_1.1 | Out-of-fold mean of the response (class 1) grouped by: ['diag\_1'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.4637 |
| 12 | 95\_TruncSVD: discharge\_disposition\_id: num\_procedures: number\_inpatient: patient\_nbr.1 | Component #2 of truncated SVD of ['discharge\_disposition\_id', 'num\_procedures', 'number\_inpatient', 'patient\_nbr'] into 3 components | Truncated SVD | 0.454 |
| 13 | 142\_NumToCatTE: admission\_source\_id: discharge\_disposition\_id: number\_inpatient: patient\_nbr.1 | Out-of-fold mean of the response (class 1) grouped by: ['admission\_source\_id', 'discharge\_disposition\_id', 'number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 10 equally populated bins) [internal parameters:(20, 3, 10)] | Numeric to Categorical Target Encoding | 0.4533 |
| 14 | 84\_ClusterTE: ClusterID10: number\_inpatient.0 | Out-of-fold mean of the response (class 0) grouped by: ['ClusterID10:number\_inpatient'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.4405 |
| 15 | 113\_NumCatTE: medical\_specialty: number\_inpatient: patient\_nbr: payer\_code.2 | Out-of-fold mean of the response (class 2) grouped by: ['medical\_specialty', 'number\_inpatient', 'patient\_nbr', 'payer\_code'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(20, 3, 10)] | Cross Validation Target Encoding | 0.3427 |
| 16 | 75\_NumToCatTE: discharge\_disposition\_id.0 | Out-of-fold mean of the response (class 0) grouped by: ['discharge\_disposition\_id'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(10, 5, 10)] | Numeric to Categorical Target Encoding | 0.3393 |
| 17 | 113\_NumCatTE: medical\_specialty: number\_inpatient: patient\_nbr: payer\_code.1 | Out-of-fold mean of the response (class 1) grouped by: ['medical\_specialty', 'number\_inpatient', 'patient\_nbr', 'payer\_code'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(20, 3, 10)] | Cross Validation Target Encoding | 0.3153 |
| 18 | 49\_NumToCatTE: number\_diagnoses: number\_inpatient.1 | Out-of-fold mean of the response (class 1) grouped by: ['number\_diagnoses', 'number\_inpatient'] using 5 folds (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(20, 5, 20)] | Numeric to Categorical Target Encoding | 0.3081 |
| 19 | 75\_NumToCatTE: discharge\_disposition\_id.1 | Out-of-fold mean of the response (class 1) grouped by: ['discharge\_disposition\_id'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(10, 5, 10)] | Numeric to Categorical Target Encoding | 0.295 |
| 20 | 103\_ClusterTE: ClusterID10: discharge\_disposition\_id.2 | Out-of-fold mean of the response (class 2) grouped by: ['ClusterID10:discharge\_disposition\_id'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.2663 |
| 21 | 142\_NumToCatTE: admission\_source\_id: discharge\_disposition\_id: number\_inpatient: patient\_nbr.0 | Out-of-fold mean of the response (class 0) grouped by: ['admission\_source\_id', 'discharge\_disposition\_id', 'number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 10 equally populated bins) [internal parameters:(20, 3, 10)] | Numeric to Categorical Target Encoding | 0.2637 |
| 22 | 39\_CVTE: diag\_2.2 | Out-of-fold mean of the response (class 2) grouped by: ['diag\_2'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.2403 |
| 23 | 12\_CVTE: diag\_3.2 | Out-of-fold mean of the response (class 2) grouped by: ['diag\_3'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.226 |
| 24 | 29\_CVTE: diag\_1.0 | Out-of-fold mean of the response (class 0) grouped by: ['diag\_1'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.2047 |
| 25 | 117\_ClusterDist50: num\_medications: number\_inpatient.0 | Distances to cluster center after segmenting columns ['num\_medications', 'number\_inpatient'] into 50 clusters. Distance to cluster #0 | Cluster Distance | 0.1976 |
| 26 | 65\_NumToCatTE: discharge\_disposition\_id: number\_emergency: number\_inpatient: patient\_nbr.2 | Out-of-fold mean of the response (class 2) grouped by: ['discharge\_disposition\_id', 'number\_emergency', 'number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 3, None)] | Numeric to Categorical Target Encoding | 0.1969 |
| 27 | 84\_ClusterTE: ClusterID10: number\_inpatient.2 | Out-of-fold mean of the response (class 2) grouped by: ['ClusterID10:number\_inpatient'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.1944 |
| 28 | 39\_CVTE: diag\_2.1 | Out-of-fold mean of the response (class 1) grouped by: ['diag\_2'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.1905 |
| 29 | 65\_NumToCatTE: discharge\_disposition\_id: number\_emergency: number\_inpatient: patient\_nbr.1 | Out-of-fold mean of the response (class 1) grouped by: ['discharge\_disposition\_id', 'number\_emergency', 'number\_inpatient', 'patient\_nbr'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 3, None)] | Numeric to Categorical Target Encoding | 0.1857 |
| 30 | 75\_NumToCatTE: discharge\_disposition\_id.2 | Out-of-fold mean of the response (class 2) grouped by: ['discharge\_disposition\_id'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(10, 5, 10)] | Numeric to Categorical Target Encoding | 0.1782 |
| 31 | 152\_NumCatTE: admission\_source\_id: glimepiride-pioglitazone: number\_diagnoses: pioglitazone.1 | Out-of-fold mean of the response (class 1) grouped by: ['admission\_source\_id', 'glimepiride-pioglitazone', 'number\_diagnoses', 'pioglitazone'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(10, 10, 100)] | Cross Validation Target Encoding | 0.1746 |
| 32 | 146\_NumToCatTE: discharge\_disposition\_id: number\_outpatient.0 | Out-of-fold mean of the response (class 0) grouped by: ['discharge\_disposition\_id', 'number\_outpatient'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 1, 100)] | Numeric to Categorical Target Encoding | 0.165 |
| 33 | 97\_ClusterDist20: number\_inpatient.12 | Distances to cluster center after segmenting columns ['number\_inpatient'] into 20 clusters. Distance to cluster #12 | Cluster Distance | 0.151 |
| 34 | 49\_NumToCatTE: number\_diagnoses: number\_inpatient.2 | Out-of-fold mean of the response (class 2) grouped by: ['number\_diagnoses', 'number\_inpatient'] using 5 folds (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(20, 5, 20)] | Numeric to Categorical Target Encoding | 0.147 |
| 35 | 12\_CVTE: diag\_3.1 | Out-of-fold mean of the response (class 1) grouped by: ['diag\_3'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.1462 |
| 36 | 152\_NumCatTE: admission\_source\_id: glimepiride-pioglitazone: number\_diagnoses: pioglitazone.2 | Out-of-fold mean of the response (class 2) grouped by: ['admission\_source\_id', 'glimepiride-pioglitazone', 'number\_diagnoses', 'pioglitazone'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(10, 10, 100)] | Cross Validation Target Encoding | 0.1412 |
| 37 | 112\_CVTE: diag\_2: glyburide-metformin.1 | Out-of-fold mean of the response (class 1) grouped by: ['diag\_2', 'glyburide-metformin'] using 5 folds [internal parameters:(100, 3, None)] | Cross Validation Target Encoding | 0.1363 |
| 38 | 45\_CVTE: diabetesMed.0 | Out-of-fold mean of the response (class 0) grouped by: ['diabetesMed'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.1355 |
| 39 | 193\_ClusterID10: number\_emergency | Centroid assignment after segmenting columns ['number\_emergency'] into 10 clusters | 193\_ClusterID10: number\_emergency | 0.1323 |
| 40 | 146\_NumToCatTE: discharge\_disposition\_id: number\_outpatient.1 | Out-of-fold mean of the response (class 1) grouped by: ['discharge\_disposition\_id', 'number\_outpatient'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 1, 100)] | Numeric to Categorical Target Encoding | 0.1306 |
| 41 | 100\_CVCatNumEnc: diag\_3: glyburide-metformin: miglitol: payer\_code: discharge\_disposition\_id.mean | Out-of-fold mean of 'discharge\_disposition\_id' grouped by: ['diag\_3', 'glyburide-metformin', 'miglitol', 'payer\_code'] using 5 folds [internal parameters:('mean', 100)] | Cross Validation Categorical to Numeric Encoding | 0.1291 |
| 42 | 74\_CVCatNumEnc: metformin-rosiglitazone: payer\_code: patient\_nbr.mean | Out-of-fold mean of 'patient\_nbr' grouped by: ['metformin-rosiglitazone', 'payer\_code'] using 5 folds [internal parameters:('mean', 10)] | Cross Validation Categorical to Numeric Encoding | 0.1267 |
| 43 | 28\_CVTE: age.1 | Out-of-fold mean of the response (class 1) grouped by: ['age'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.1213 |
| 44 | 112\_CVTE: diag\_2: glyburide-metformin.2 | Out-of-fold mean of the response (class 2) grouped by: ['diag\_2', 'glyburide-metformin'] using 5 folds [internal parameters:(100, 3, None)] | Cross Validation Target Encoding | 0.1212 |
| 45 | 83\_ClusterTE: ClusterID20: discharge\_disposition\_id: number\_inpatient: patient\_nbr.0 | Out-of-fold mean of the response (class 0) grouped by: ['ClusterID20:discharge\_disposition\_id:number\_inpatient:patient\_nbr'] using 3 folds (Clustered into 20 clusters) [internal parameters:(20, 10, 3, 10)] | Cluster Target Encoding | 0.1181 |
| 46 | 84\_ClusterTE: ClusterID10: number\_inpatient.1 | Out-of-fold mean of the response (class 1) grouped by: ['ClusterID10:number\_inpatient'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.117 |
| 47 | 80\_ClusterTE: ClusterID10: discharge\_disposition\_id: number\_inpatient: patient\_nbr.2 | Out-of-fold mean of the response (class 2) grouped by: ['ClusterID10:discharge\_disposition\_id:number\_inpatient:patient\_nbr'] using 3 folds (Clustered into 10 clusters) [internal parameters:(10, 10, 3, 10)] | Cluster Target Encoding | 0.113 |
| 48 | 12\_CVTE: diag\_3.0 | Out-of-fold mean of the response (class 0) grouped by: ['diag\_3'] using 5 folds [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.1115 |
| 49 | 96\_NumCatTE: diag\_2: discharge\_disposition\_id: metformin-pioglitazone: metformin-rosiglitazone: number\_inpatient: payer\_code.1 | Out-of-fold mean of the response (class 1) grouped by: ['diag\_2', 'discharge\_disposition\_id', 'metformin-pioglitazone', 'metformin-rosiglitazone', 'number\_inpatient', 'payer\_code'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(20, 5, 100)] | Cross Validation Target Encoding | 0.1099 |
| 50 | 146\_NumToCatTE: discharge\_disposition\_id: number\_outpatient.2 | Out-of-fold mean of the response (class 2) grouped by: ['discharge\_disposition\_id', 'number\_outpatient'] using 5 folds (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 1, 100)] | Numeric to Categorical Target Encoding | 0.1082 |

## Final Model

**Pipeline**

Final StackedEnsemble pipeline with ensemble\_level=1 transforming 37 original features -> 342 features in each of 4 models each fit on 4 internal holdout splits then linearly blended

**Details**

* The fitted features of the final model are the best features found during the feature engineering iterations.
* The target transformer indicates the type of transformation applied to the target column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | XGBoostGBMModel | 1.0 | 4 | 341 | LabelEncoder |

* Model Index: 0 has a weight of 1.0 in the final ensemble

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **grow policy** | **tree method** | **subsample** | **colsample bytree** | **model class name** | **learning rate** | **max leaves** | **max depth** | **index** |
| XGBoostGBMModel | depthwise | gpu\_hist | 0.8 | 0.45 | XGBoostGBMModel | 0.05 | 0 | 8 | 0 |

For a complete list of the parameters of the final model, see the Appendix.

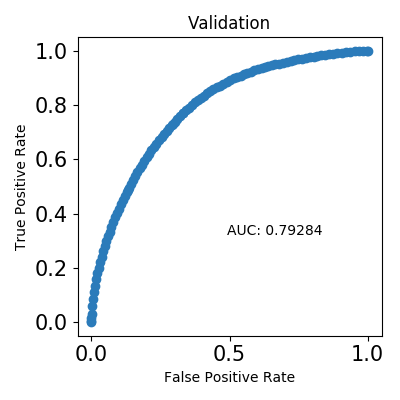
**Performance of Final Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scorer** | **Optimized** | **Better score is** | **Final ensemble scores on validation (internal or external holdout(s)) data** | **Final ensemble standard deviation on validation (internal or external holdout(s)) data** |
| ACCURACY |  | higher | 0.60277 | 0.0029973 |
| AUC | \* | higher | 0.79284 | 0.0021647 |
| AUCPR |  | higher | 0.65647 | 0.0034799 |
| F05 |  | higher | 0.60277 | 0.0029973 |
| F1 |  | higher | 0.60277 | 0.0029973 |
| F2 |  | higher | 0.71429 | 2.2204e-16 |
| GINI |  | higher | 0.58569 | 0.0043294 |
| LOGLOSS |  | lower | 0.84793 | 0.0038709 |
| MACROAUC |  | higher | 0.70075 | 0.0026703 |
| MCC |  | higher | 0.40415 | 0.0044959 |

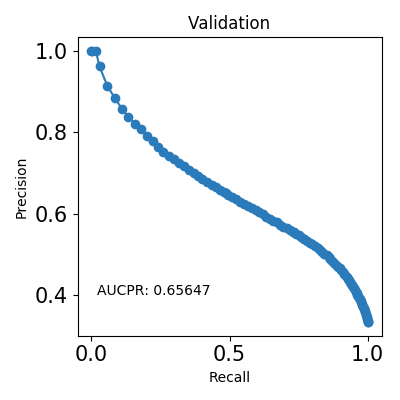
**Validation Confusion Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Predicted: <30** | **Predicted: >30** | **Predicted: NO** | **error** |
| Actual: <30 | 437 | 4,911 | 6,009 | 96% |
| Actual: >30 | 271 | 16,050 | 19,224 | 55% |
| Actual: NO | 153 | 9,871 | 44,840 | 18% |

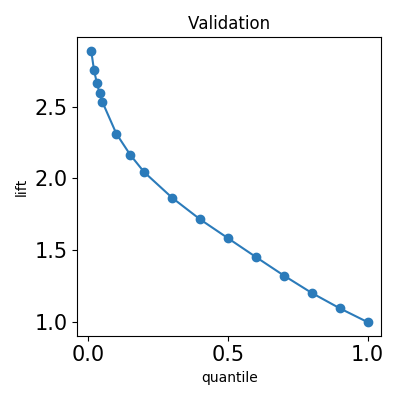
*Receiver Operating Characteristic Curve*



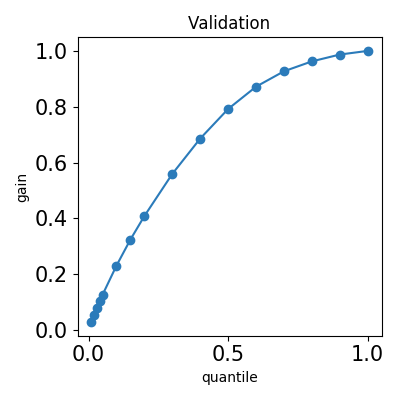
*Precision Recall Curve*



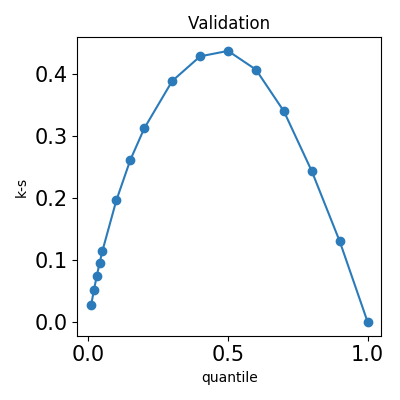
*Cumulative Lift*



*Cumulative Gains*



*Kolmogorov–Smirnov*



## Alternative Models

During the experiment, Driverless AI trained 73 alternative models. The following algorithms were evaluated during the Driverless AI experiment:

|  |  |  |  |
| --- | --- | --- | --- |
| **algorithm** | **package** | **version** | **documentation** |
| gbtree | xgboost | 0.90 | XGBoost: eXtreme Gradient Boosting library. Contributors: https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md |
| gblinear | xgboost | 0.90 | XGBoost: eXtreme Gradient Boosting library. Contributors: https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md |
| lightgbm | lightgbm | 2.2.4 | LightGBM, Light Gradient Boosting Machine. Contributors: https://github.com/microsoft/LightGBM/graphs/contributors. |

Driverless AI is able to evaluate the algorithms: XGBoost GBM, XGBoost Dart, XGBoost GLM, LightGBM, RuleFit, Tensorflow, and FTRL models. The table below explains why certain algorithms were not selected for the final model, if any.

|  |  |
| --- | --- |
| **algorithm** | **selection** |
| rulefit | algorithm not evaluated due to experiment configuration |
| tensorflow | algorithm not evaluated due to experiment configuration |
| ftrl | algorithm not evaluated due to experiment configuration |
| dart | algorithm not evaluated due to experiment configuration |
| gblinear | not selected due to low performance during model tuning stage |
| lightgbm | not selected due to low performance during feature evolution stage |
| gbtree | selected for final model |

## 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\_825d312a-1d57-11ea-b321-0242ac110002/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, 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).

## Partial Dependence Plots

Partial dependence plots are currently not supported for multiclass classification.

## Appendix

### Final Model Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | XGBoostGBMModel | 1.0 | 4 | 341 | LabelEncoder |

**Model Index: 0**

|  |  |
| --- | --- |
| **parameter** | **value** |
| accuracy | 5 |
| booster | gbtree |
| colsample\_bytree | 0.45 |
| debug\_verbose | 0 |
| disable\_gpus | False |
| dummy | False |
| early\_stopping\_rounds | 100 |
| early\_stopping\_threshold | 0 |
| encoder |  |
| ensemble\_level | 1 |
| eval\_metric | mlogloss |
| gamma | 0.01 |
| gpu\_id | 0 |
| grow\_policy | depthwise |
| interpretability | 5 |
| labels | ['<30', '>30', 'NO'] |
| learning\_rate | 0.05 |
| lossguide | False |
| max\_bin | 256 |
| max\_delta\_step | 0.0 |
| max\_depth | 8 |
| max\_leaves | 0 |
| min\_child\_weight | 1 |
| model\_class\_name | XGBoostGBMModel |
| model\_id | 0 |
| model\_origin | SEQUENCE |
| monotonicity\_constraints | False |
| n\_estimators | 500 |
| n\_gpus | 1 |
| n\_jobs | 3 |
| num\_class | 3 |
| num\_classes | 3 |
| objective | multi:softprob |
| pred\_gap |  |
| pred\_periods |  |
| random\_state | 822249717 |
| reg\_alpha | 0.0 |
| reg\_lambda | 5.0 |
| scale\_pos\_weight | 1.0 |
| score\_f\_name | AUC |
| seed | 822249717 |
| silent | True |
| subsample | 0.8 |
| target |  |
| tgc |  |
| time\_column |  |
| time\_tolerance | 5 |
| train\_shape | [101766, 48] |
| tree\_method | gpu\_hist |
| tsp |  |
| valid\_shape |  |
| nfolds | 4 |

**Config Overrides**

The Config Overrides represent the fine-control parameters.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| enable\_benchmark | False |
| vis\_server\_port | 12346 |
| procsy\_port | 12347 |
| h2o\_port | 12348 |
| master\_redis\_password | t1Oi8w3coZczF50428ABc6XCCsBvlU3O |
| data\_directory | ./tmp |
| enable\_startup\_checks | False |
| authentication\_method | local |
| local\_htpasswd\_file | /config/htpasswd |
| prob\_lagsinteraction | 0.2 |
| prob\_lagsaggregates | 0.2 |
| prob\_default\_lags | 0.2 |
| prob\_lag\_non\_targets | 0.1 |
| included\_scorers | ['ACCURACY', 'AUC', 'AUCPR', 'F05', 'F1', 'F2', 'GINI', 'LOGLOSS', 'MACROAUC', 'MAE', 'MAPE', 'MCC', 'MER', 'MSE', 'R2', 'RMSE', 'RMSLE', 'RMSPE', 'SMAPE'] |
| included\_models | ['FTRL', 'GLM', 'IMBALANCEDLIGHTGBM', 'IMBALANCEDXGBOOSTGBM', 'LIGHTGBM', 'RULEFIT', 'TENSORFLOW', 'XGBOOSTDART', 'XGBOOSTGBM'] |
| included\_transformers | ['CVCatNumEncodeTransformer', 'CVTargetEncodeTransformer', 'CatOriginalTransformer', 'CatTransformer', 'ClusterDistTransformer', 'ClusterIdTransformer', 'ClusterTETransformer', 'DateOriginalTransformer', 'DateTimeOriginalTransformer', 'DatesTransformer', 'EwmaLagsTransformer', 'FrequentTransformer', 'InteractionsTransformer', 'IsHolidayTransformer', 'IsolationForestAnomalyNumCatAllColsTransformer', 'IsolationForestAnomalyNumCatTransformer', 'IsolationForestAnomalyNumericTransformer', 'LagsAggregatesTransformer', 'LagsInteractionTransformer', 'LagsTransformer', 'LexiLabelEncoderTransformer', 'NumCatTETransformer', 'NumToCatTETransformer', 'NumToCatWoEMonotonicTransformer', 'NumToCatWoETransformer', 'OneHotEncodingTransformer', 'OriginalTransformer', 'TextBiGRUTransformer', 'TextCNNTransformer', 'TextCharCNNTransformer', 'TextLinModelTransformer', 'TextTransformer', 'TruncSVDNumTransformer', 'WeightOfEvidenceTransformer'] |
| override\_lag\_sizes |  |
| experiment\_id | 825d312a-1d57-11ea-b321-0242ac110002 |
| experiment\_tmp\_dir | ./tmp/h2oai\_experiment\_825d312a-1d57-11ea-b321-0242ac110002 |