**Driverless AI Experiment: *huhacoho***

Generated by: training

Generated on: Fri Sep 13 04:12:31 2019

[Experiment Overview 1](#_Toc7449951)

[Data Overview 3](#_Toc7449952)

[Methodology 5](#_Toc7449953)

[Data Sampling 10](#_Toc7449954)

[Validation Strategy 10](#_Toc7449955)

[Model Tuning 12](#_Toc7449956)

[Feature Evolution 13](#_Toc7449957)

[Feature Transformations 13](#_Toc7449958)

[Final Model 14](#_Toc7449959)

[Alternative Models 17](#_Toc7449960)

[Deployment 18](#_Toc7449961)

[Partial Dependence Plots 19](#_Toc7449962)

[Appendix 19](#_Toc7449963)

## Experiment Overview

Driverless AI built a stacked ensemble of 1 LightGBMModel to predict *Default* given 23 original features from the input dataset *CreditCard\_train*. This classification experiment completed in 10 minutes and 42 seconds (0:10:42), using 18 of the 23 original features, and 45 of the 290 engineered features.

### Performance

|  |  |
| --- | --- |
| **Dataset** | **AUC** |
| Internal Validation | 0.778 |
| Test Data | 0.782 |

### 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 | 2 | 1-10 |
| Interpretability | Controls complexity of the model | 6 | 1-10 |

### System Specifications

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Address** | **System** | **System Memory** | **CPUs** | **GPUs** |
| http://127.0.0.1:12345 | Docker/Linux | 60 | 4 | 1 |

### Versions

|  |
| --- |
| Driverless AI Version |
| 1.7.0 |

## Data Overview

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

|  |  |  |  |
| --- | --- | --- | --- |
| **data** | **file path** | **number of rows** | **number of columns** |
| training | ./tmp/c94cc0c0-aebf-11e9-9d83-0242ac110002/CreditCard\_train.1564047435.1649468.bin | 23,999 | 25 |
| validation | Not provided | None | None |
| testing | ./tmp/c953ca46-aebf-11e9-9d83-0242ac110002/CreditCard\_test.1564047435.1719637.bin | 6,000 | 25 |

### 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** |
| ID | int | 1.000 | 15,020.374 | 30,000.000 | 8,662.213 | 23,999 | 1 |
| CreditLimit | int | 10,000.000 | 167,221.121 | 1,000,000.000 | 129,787.817 | 81 | 2,716 |
| Education | int | 0.000 | 1.853 | 6.000 | 0.792 | 7 | 11,198 |
| Age | int | 21.000 | 35.415 | 79.000 | 9.193 | 56 | 1,265 |
| Status1 | int | -2.000 | -0.148 | 1.000 | 0.875 | 4 | 11,804 |
| Status2 | int | -2.000 | -0.157 | 2.000 | 1.127 | 5 | 12,617 |
| Status3 | int | -2.000 | -0.175 | 3.000 | 1.158 | 6 | 12,618 |
| Status4 | int | -2.000 | -0.227 | 4.000 | 1.131 | 7 | 13,146 |
| Status5 | int | -2.000 | -0.271 | 5.000 | 1.108 | 7 | 13,551 |
| Status6 | int | -2.000 | -0.294 | 6.000 | 1.138 | 8 | 13,056 |
| BillAmt1 | int | -165,580.000 | 51,009.063 | 964,511.000 | 73,011.145 | 18,729 | 1,557 |
| BillAmt2 | int | -67,526.000 | 48,926.229 | 983,931.000 | 70,352.575 | 18,396 | 1,982 |
| BillAmt3 | int | -157,264.000 | 46,754.180 | 1,664,089.000 | 68,844.630 | 18,165 | 2,253 |
| BillAmt4 | int | -170,000.000 | 43,052.092 | 891,586.000 | 63,804.779 | 17,792 | 2,527 |
| BillAmt5 | int | -61,372.000 | 40,098.418 | 927,171.000 | 60,375.981 | 17,337 | 2,777 |
| BillAmt6 | int | -339,603.000 | 38,674.051 | 961,664.000 | 59,221.131 | 16,996 | 3,205 |
| PayAmt1 | int | 0.000 | 5,586.367 | 873,552.000 | 15,873.868 | 6,907 | 4,158 |
| PayAmt2 | int | 0.000 | 5,865.566 | 1,684,259.000 | 22,767.090 | 6,895 | 4,257 |
| PayAmt3 | int | 0.000 | 5,244.086 | 896,040.000 | 17,890.099 | 6,578 | 4,724 |
| PayAmt4 | int | 0.000 | 4,848.374 | 621,000.000 | 15,833.880 | 6,026 | 5,058 |
| PayAmt5 | int | 0.000 | 4,744.817 | 426,529.000 | 14,960.190 | 6,000 | 5,344 |
| PayAmt6 | int | 0.000 | 5,105.771 | 528,666.000 | 17,089.146 | 6,004 | 5,722 |

#### Boolean Columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **name** | **data\_type** | **min** | **mean** | **max** | **std** | **freq of max value** |
| Default | bool | False | 0.2212 | True | 0.415 | 5,308 |

#### Categorical Columns

|  |  |  |  |
| --- | --- | --- | --- |
| **name** | **unique** | **top** | **freq of top value** |
| Sex | 2 | F | 9,434 |
| Marriage | 4 | S | 12,808 |

### 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 checked the train and test data for any shift in distributions but found none. This indicates that all the predictors/columns in the train and test data are from the same distribution.

## Methodology

This section describes the experiment methodology.

### Assumptions and Limitations

Driverless AI trains all models based on the training data provided (in this case: *CreditCard\_train*). 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 and test data. 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 performed shift detection but found no significant changes in the distribution of the train and test data.

### 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 linear, xgboost 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
  + 16 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 **290** features over **20** iterations
  + 6 models trained and scored to further evaluate engineered features
* **Final Model** 
  + the final model is a stacked ensemble of **1 LightGBMModel**
  + the features of this model are the best features found during the feature engineering iterations
* **Create Scoring Pipeline** 
  + created and exported the MOJO and Python scoring pipeline
  + MOJO Scoring Pipeline: h2oai\_experiment\_854f239a-d5d9-11e9-b3a7-0242ac110002/mojo\_pipeline/mojo.zip
  + Python Scoring Pipeline: h2oai\_experiment\_854f239a-d5d9-11e9-b3a7-0242ac110002/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 | 27.02 | 0 |
| Model and Feature Tuning | 157.88 | 16 |
| Feature Evolution | 40.43 | 6 |
| Final Pipeline Training | 287.28 | 4 |

### Experiment Settings

Below are the settings selected for the experiment by training. The Defined Parameters represent the high-level parameters; the Config Overrides represent the fine-control parameters. Note: the settings listed below do not differentiate between what a user explicitly set and what DAI automatically sets.

**Defined Parameters**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| is\_classification | True |
| enable\_gpus | True |
| seed | True |
| accuracy | 5 |
| time | 2 |
| interpretability | 6 |
| time\_groups\_columns | None |
| num\_prediction\_periods | None |
| num\_gap\_periods | None |
| is\_timeseries | False |

**Config Overrides**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| config\_overrides |  |
| max\_runtime\_minutes | 0 |
| recipe | auto |
| feature\_engineering\_effort | 5 |
| check\_distribution\_shift | auto |
| check\_distribution\_shift\_drop | auto |
| drop\_features\_distribution\_shift\_threshold\_auc | 0.6 |
| check\_leakage | auto |
| drop\_features\_leakage\_threshold\_auc | 0.999 |
| make\_python\_scoring\_pipeline | True |
| make\_mojo\_scoring\_pipeline | False |
| max\_cores | -1 |
| num\_gpus\_per\_experiment | -1 |
| num\_gpus\_per\_model | 1 |
| gpu\_id\_start | 0 |
| min\_num\_rows | 100 |
| reproducibility\_level | 1 |
| seed | 98765 |
| feature\_evolution\_data\_size | 100000000 |
| max\_orig\_cols\_selected | 10000 |
| max\_relative\_cardinality | 0.95 |
| allow\_different\_classes\_across\_fold\_splits | True |
| fixed\_ensemble\_level | -1 |
| enable\_imbalanced\_sampling | auto |
| quantile\_imbalanced\_sampling | False |
| max\_num\_classes | 200 |
| feature\_brain\_level | 0 |
| feature\_brain\_save\_every\_iteration | 0 |
| which\_iteration\_brain | -1 |
| min\_dai\_iterations | 0 |
| nfeatures\_max | -1 |
| tensorflow\_max\_epochs\_nlp | 2 |
| enable\_tensorflow\_nlp\_accuracy\_switch | 5 |
| enable\_tensorflow\_textcnn | auto |
| enable\_tensorflow\_textbigru | auto |
| enable\_tensorflow\_charcnn | auto |
| tensorflow\_nlp\_pretrained\_embeddings\_file\_path |  |
| text\_fraction\_for\_text\_dominated\_problem | 0.3 |
| text\_transformer\_fraction\_for\_text\_dominated\_problem | 0.3 |
| string\_col\_as\_text\_threshold | 0.3 |
| max\_feature\_interaction\_depth | -1 |
| parameter\_tuning\_num\_models | -1 |
| target\_transformer | auto |
| fixed\_num\_folds | 0 |
| enable\_target\_encoding | True |
| included\_transformers | ['CVCatNumEncodeTransformer', 'CVTargetEncodeTransformer', 'CatOriginalTransformer', 'ClusterDistTransformer', 'ClusterIdTransformer', 'ClusterTETransformer', 'DateOriginalTransformer', 'DateTimeOriginalTransformer', 'DatesTransformer', 'EwmaLagsTransformer', 'FrequentTransformer', 'InteractionsTransformer', 'IsHolidayTransformer', 'LagsAggregatesTransformer', 'LagsInteractionTransformer', 'LagsTransformer', 'NumCatTETransformer', 'NumToCatTETransformer', 'NumToCatWoEMonotonicTransformer', 'NumToCatWoETransformer', 'OneHotEncodingTransformer', 'OriginalTransformer', 'TextBiGRUTransformer', 'TextCNNTransformer', 'TextCharCNNTransformer', 'TextLinModelTransformer', 'TextTransformer', 'TruncSVDNumTransformer', 'WeightOfEvidenceTransformer'] |
| included\_models | ['FTRL', 'GLM', 'LIGHTGBM', 'RULEFIT', 'TENSORFLOW', 'XGBOOSTDART', 'XGBOOSTGBM'] |
| included\_scorers | ['ACCURACY', 'AUC', 'AUCPR', 'F05', 'F1', 'F2', 'GINI', 'LOGLOSS', 'MACROAUC', 'MAE', 'MAPE', 'MCC', 'MER', 'MSE', 'R2', 'RMSE', 'RMSLE', 'RMSPE', 'SMAPE'] |
| enable\_xgboost\_gbm | auto |
| enable\_xgboost\_dart | auto |
| enable\_glm | auto |
| enable\_lightgbm | auto |
| enable\_rf | auto |
| enable\_tensorflow | off |
| enable\_rulefit | auto |
| enable\_ftrl | auto |
| max\_nestimators | 3000 |
| max\_nestimators\_feature\_evolution\_factor | 0.2 |
| max\_learning\_rate | 0.5 |
| max\_epochs | 10 |
| rulefit\_max\_num\_rules | -1 |
| drop\_constant\_columns | True |
| detailed\_traces | False |
| time\_series\_recipe | True |
| holiday\_features | True |
| override\_lag\_sizes |  |
| min\_lag\_size | -1 |
| allow\_tgc\_memorization | False |
| tgc\_only\_use\_all\_groups | True |
| time\_series\_holdout\_preds | True |
| time\_series\_max\_n\_splits | 20 |
| prob\_lag\_non\_targets | 0.1 |
| dump\_varimp\_every\_scored\_indiv | False |
| dump\_modelparams\_every\_scored\_indiv | True |
| skip\_transformer\_failures | True |
| skip\_model\_failures | True |
| detailed\_skip\_failure\_messages\_level | 1 |

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 | 10 |
| number of models trained per iteration | 4 |
| early stopping rounds | 5 |
| 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 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).

## 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 linear, 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 greatest score and lowest training time.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **job order** | **booster** | **nfeatures** | **scores** | **training times** |
| 12 | lightgbm | 65 | 0.7577465668 | 3.0018465519 |
| 9 | gbtree | 23 | 0.7572478141 | 1.9028661251 |
| 11 | gblinear | 48 | 0.7414762432 | 2.6672370434 |
| 10 | gblinear | 45 | 0.7344633872 | 1.3046066761 |
| 7 | gblinear | 23 | 0.6858783913 | 13.2266044617 |
| 4 | lightgbm | 22 | 0.7568998012 | 0.9699571133 |
| 0 | lightgbm | 21 | 0.7568371427 | 0.9792561531 |
| 3 | lightgbm | 23 | 0.7557175993 | 1.9054424763 |
| 14 | lightgbm | 68 | 0.7556708861 | 1.8950619698 |
| 15 | gbtree | 61 | 0.7531963919 | 2.3083953857 |

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

### gblinear tuning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **reg alpha** | **reg lambda** | **n lambda** | **min lambda fraction** | **nfeatures** | **scores** | **training times** |
| 0.0005 | 0.001 | 1 | 0.0001 | 48 | 0.7414762432 | 2.6672370434 |
| 0.0005 | 0.001 | 1 | 0.0001 | 45 | 0.7344633872 | 1.3046066761 |
| 0.0005 | 0.001 | 1 | 0.0001 | 23 | 0.6858783913 | 13.2266044617 |
| 0.0005 | 0.001 | 1 | 0.0001 | 53 | 0.7475268533 | 6.6992106438 |

### 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 | 23 | 0.7572478141 | 1.9028661251 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 61 | 0.7531963919 | 2.3083953857 |

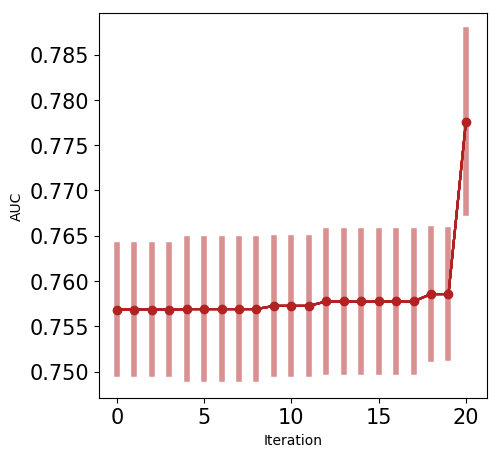
### lightgbm tuning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | 65 | 0.7577465668 | 3.0018465519 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 22 | 0.7568998012 | 0.9699571133 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 21 | 0.7568371427 | 0.9792561531 |
| gpu\_hist | depthwise | 10.0 | 0.0 | 0.8 | 0.7 | 23 | 0.7557175993 | 1.9054424763 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 68 | 0.7556708861 | 1.8950619698 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.8 | 0.7 | 20 | 0.7521109401 | 0.9082331657 |
| gpu\_hist | depthwise | 10.0 | 0.0 | 0.8 | 0.7 | 21 | 0.7503641829 | 2.0553398132 |

## 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 (6) 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.



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

* **OriginalTransformer**: applies an identity transformation to a numeric column.
* **CatOriginalTransformer**: applies an identity transformation that leaves categorical features as they are. This transformer works with models that can handle non-numeric feature values.
* **DateTimeOriginalTransformer**: None
* **DateOriginalTransformer**: None

## 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 , restricted to those with relative importance greater than or equal to . If no transformer was applied, the feature is an original column.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Feature** | **Description** | **Transformer** | **Relative Importance** |
| 1 | 45\_NumToCatWoE: PayAmt3: Status1: Status2.0 | Weight of Evidence for columns ['PayAmt3', 'Status1', 'Status2'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 1.0 |
| 2 | 61\_NumToCatWoE: PayAmt2: Status1: Status2.0 | Weight of Evidence for columns ['PayAmt2', 'Status1', 'Status2'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.6885 |
| 3 | 51\_ClusterTE: ClusterID71: PayAmt3: Status1: Status3.0 | Out-of-fold mean of the response grouped by: ['ClusterID71:PayAmt3:Status1:Status3'] using 5 folds (Clustered into 71 clusters) [internal parameters:(71, True, 20, 3, 100)] | Cluster Target Encoding | 0.2497 |
| 4 | 53\_ClusterTE: ClusterID60: CreditLimit: PayAmt5: Status1.0 | Out-of-fold mean of the response grouped by: ['ClusterID60:CreditLimit:PayAmt5:Status1'] using 5 folds (Clustered into 60 clusters) [internal parameters:(60, True, 20, 5, None)] | Cluster Target Encoding | 0.1295 |
| 5 | 73\_WoE: CreditLimit: Status1: Status2.0 | Weight of Evidence for columns ['CreditLimit', 'Status1', 'Status2'] column #0 | Weight of Evidence | 0.0978 |
| 6 | 44\_CVCatNumEnc: Status1: Status2: Status3: Status6: CreditLimit.mean | Out-of-fold mean of 'CreditLimit' grouped by: ['Status1', 'Status2', 'Status3', 'Status6'] using 5 folds [internal parameters:('mean', 10)] | Cross Validation Categorical to Numeric Encoding | 0.0973 |
| 7 | 56\_NumToCatWoE: CreditLimit: PayAmt1: Status5.0 | Weight of Evidence for columns ['CreditLimit', 'PayAmt1', 'Status5'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0907 |
| 8 | 72\_WoE: CreditLimit: Status1: Status2: Status4.0 | Weight of Evidence for columns ['CreditLimit', 'Status1', 'Status2', 'Status4'] column #0 | Weight of Evidence | 0.0649 |
| 9 | 33\_NumToCatWoE: BillAmt5: Status5.0 | Weight of Evidence for columns ['BillAmt5', 'Status5'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0645 |
| 10 | 48\_NumCatTE: BillAmt2: CreditLimit: PayAmt4: PayAmt6: Status1: Status3.0 | Out-of-fold mean of the response grouped by: ['BillAmt2', 'CreditLimit', 'PayAmt4', 'PayAmt6', 'Status1', 'Status3'] using 5 folds (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(20, 10, 20)] | Cross Validation Target Encoding | 0.0638 |
| 11 | 10\_PayAmt2 | PayAmt2 (original) | None | 0.0606 |
| 12 | 47\_WoE: Age: Status1: Status4: Status5.0 | Weight of Evidence for columns ['Age', 'Status1', 'Status4', 'Status5'] column #0 | Weight of Evidence | 0.0573 |
| 13 | 68\_NumToCatWoE: BillAmt6: PayAmt3.0 | Weight of Evidence for columns ['BillAmt6', 'PayAmt3'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0572 |
| 14 | 39\_NumCatTE: CreditLimit: Status1: Status2: Status5.0 | Out-of-fold mean of the response grouped by: ['CreditLimit', 'Status1', 'Status2', 'Status5'] using 5 folds (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(10, 3, 10)] | Cross Validation Target Encoding | 0.0569 |
| 15 | 69\_NumToCatWoE: BillAmt5: PayAmt1: PayAmt3: Status2.0 | Weight of Evidence for columns ['BillAmt5', 'PayAmt1', 'PayAmt3', 'Status2'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0565 |
| 16 | 9\_PayAmt1 | PayAmt1 (original) | None | 0.0554 |
| 17 | 7\_CreditLimit | CreditLimit (original) | None | 0.0552 |
| 18 | 64\_InteractionSub: BillAmt1: Status1 | [BillAmt1] - [Status1] | Interaction | 0.052 |
| 19 | 59\_ClusterTE: ClusterID12: BillAmt5: Status4: Status5.0 | Out-of-fold mean of the response grouped by: ['ClusterID12:BillAmt5:Status4:Status5'] using 5 folds (Clustered into 12 clusters) [internal parameters:(12, False, 10, 3, 10)] | Cluster Target Encoding | 0.0511 |
| 20 | 50\_NumCatTE: PayAmt4: PayAmt5: PayAmt6: Status1: Status2: Status3: Status4: Status5.0 | Out-of-fold mean of the response grouped by: ['PayAmt4', 'PayAmt5', 'PayAmt6', 'Status1', 'Status2', 'Status3', 'Status4', 'Status5'] using 5 folds (numeric columns are bucketed into 10 equally populated bins) [internal parameters:(10, 5, 100)] | Cross Validation Target Encoding | 0.0479 |
| 21 | 1\_BillAmt1 | BillAmt1 (original) | None | 0.0468 |
| 22 | 44\_CVCatNumEnc: Status1: Status2: Status3: Status6: PayAmt5.mean | Out-of-fold mean of 'PayAmt5' grouped by: ['Status1', 'Status2', 'Status3', 'Status6'] using 5 folds [internal parameters:('mean', 10)] | Cross Validation Categorical to Numeric Encoding | 0.0455 |
| 23 | 12\_PayAmt4 | PayAmt4 (original) | None | 0.04 |
| 24 | 11\_PayAmt3 | PayAmt3 (original) | None | 0.0385 |
| 25 | 14\_PayAmt6 | PayAmt6 (original) | None | 0.0373 |
| 26 | 57\_CVTE: CreditLimit: Status1: Status2: Status5.0 | Out-of-fold mean of the response grouped by: ['CreditLimit', 'Status1', 'Status2', 'Status5'] using 5 folds [internal parameters:(10, 10, 100)] | Cross Validation Target Encoding | 0.037 |
| 27 | 41\_CVTE: CreditLimit: Status1: Status2: Status4.0 | Out-of-fold mean of the response grouped by: ['CreditLimit', 'Status1', 'Status2', 'Status4'] using 5 folds [internal parameters:(10, 3, None)] | Cross Validation Target Encoding | 0.0368 |
| 28 | 2\_BillAmt2 | BillAmt2 (original) | None | 0.0357 |
| 29 | 78\_NumToCatWoE: PayAmt4.0 | Weight of Evidence for columns ['PayAmt4'] column #0 (numeric columns are bucketed into 100 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0316 |
| 30 | 46\_InteractionSub: Age: PayAmt5 | [Age] - [PayAmt5] | Interaction | 0.0302 |
| 31 | 38\_CVTE: CreditLimit: Status3: Status4: Status6.0 | Out-of-fold mean of the response grouped by: ['CreditLimit', 'Status3', 'Status4', 'Status6'] using 5 folds [internal parameters:(100, 5, 100)] | Cross Validation Target Encoding | 0.0295 |
| 32 | 54\_ClusterID89: BillAmt3: CreditLimit | Centroid assignment after segmenting columns ['BillAmt3', 'CreditLimit'] into 89 clusters | Transformer Unknown | 0.0292 |
| 33 | 21\_Freq: Age | Encoding of categorical levels of feature(s) ['Age'] to value between 0 and 1 based on their relative frequency | Frequency Encoding | 0.0265 |
| 34 | 52\_WoE: Status3.0 | Weight of Evidence for columns ['Status3'] column #0 | Weight of Evidence | 0.0258 |
| 35 | 3\_BillAmt3 | BillAmt3 (original) | None | 0.0252 |
| 36 | 0\_Age | Age (original) | None | 0.0251 |
| 37 | 65\_Freq: CreditLimit: Status1: Status6 | Encoding of categorical levels of feature(s) ['CreditLimit', 'Status1', 'Status6'] to their counts | Frequency Encoding | 0.0231 |
| 38 | 44\_CVCatNumEnc: Status1: Status2: Status3: Status6: BillAmt1.mean | Out-of-fold mean of 'BillAmt1' grouped by: ['Status1', 'Status2', 'Status3', 'Status6'] using 5 folds [internal parameters:('mean', 10)] | Cross Validation Categorical to Numeric Encoding | 0.0226 |
| 39 | 74\_NumCatTE: Age: PayAmt5: PayAmt6: Status1: Status4.0 | Out-of-fold mean of the response grouped by: ['Age', 'PayAmt5', 'PayAmt6', 'Status1', 'Status4'] using 5 folds (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(100, 10, 20)] | Cross Validation Target Encoding | 0.0215 |
| 40 | 20\_Status6 | Status6 (original) | None | 0.0209 |
| 41 | 4\_BillAmt4 | BillAmt4 (original) | None | 0.0207 |
| 42 | 77\_ClusterID40: Age: BillAmt6 | Centroid assignment after segmenting columns ['Age', 'BillAmt6'] into 40 clusters | Transformer Unknown | 0.0199 |
| 43 | 6\_BillAmt6 | BillAmt6 (original) | None | 0.0189 |
| 44 | 5\_BillAmt5 | BillAmt5 (original) | None | 0.0187 |
| 45 | 22\_Freq: CreditLimit | Encoding of categorical levels of feature(s) ['CreditLimit'] to value between 0 and 1 based on their relative frequency | Frequency Encoding | 0.0182 |
| 46 | 37\_Freq: Education | Encoding of categorical levels of feature(s) ['Education'] to value between 0 and 1 based on their relative frequency | Frequency Encoding | 0.0175 |
| 47 | 13\_PayAmt5 | PayAmt5 (original) | None | 0.0169 |
| 48 | 34\_Freq: Status1: Status2: Status5 | Encoding of categorical levels of feature(s) ['Status1', 'Status2', 'Status5'] to value between 0 and 1 based on their relative frequency | Frequency Encoding | 0.0157 |
| 49 | 44\_CVCatNumEnc: Status1: Status2: Status3: Status6: BillAmt2.mean | Out-of-fold mean of 'BillAmt2' grouped by: ['Status1', 'Status2', 'Status3', 'Status6'] using 5 folds [internal parameters:('mean', 10)] | Cross Validation Categorical to Numeric Encoding | 0.0148 |
| 50 | 8\_Education | Education (original) | None | 0.0145 |

## Final Model

**Pipeline**

Final StackedEnsemble pipeline with ensemble\_level=1 transforming 21 original features -> 63 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 | LightGBMModel | 1.0 | 4 | 63 | str |

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **colsample bytree** | **grow policy** | **max depth** | **learning rate** | **subsample** | **tree method** | **index** | **max leaves** |
| LightGBMModel | 0.8 | depthwise | 6 | 0.05 | 0.7 | gpu\_hist | 0 | 64 |

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** | **Final test scores** | **Final test standard deviation** |
| ACCURACY |  | higher | 0.81675 | 0.0065023 | 0.817 | 0.0065023 |
| AUC | \* | higher | 0.7776 | 0.010078 | 0.78191 | 0.010078 |
| AUCPR |  | higher | 0.5328 | 0.016822 | 0.54399 | 0.0187 |
| F05 |  | higher | 0.57051 | 0.012281 | 0.56739 | 0.016293 |
| F1 |  | higher | 0.54555 | 0.015175 | 0.55026 | 0.015175 |
| F2 |  | higher | 0.64281 | 0.012275 | 0.64046 | 0.012275 |
| GINI |  | higher | 0.5552 | 0.020156 | 0.56382 | 0.020156 |
| LOGLOSS |  | lower | 0.43409 | 0.012035 | 0.43121 | 0.012035 |
| MACROAUC |  | higher | 0.7776 | 0.010078 | 0.78191 | 0.010078 |
| MCC |  | higher | 0.4188 | 0.020102 | 0.41196 | 0.020102 |

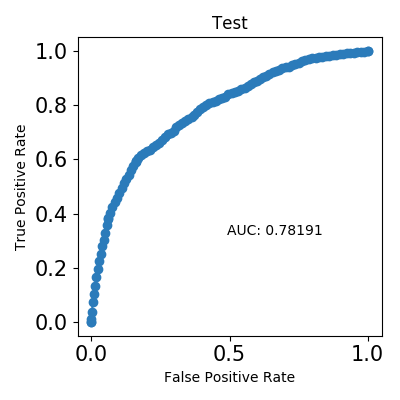
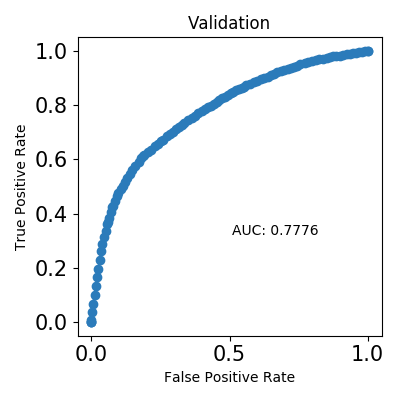
**Validation Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 15,663 | 3,028 | 16% |
| Actual: 1 | 2,233 | 3,075 | 42% |

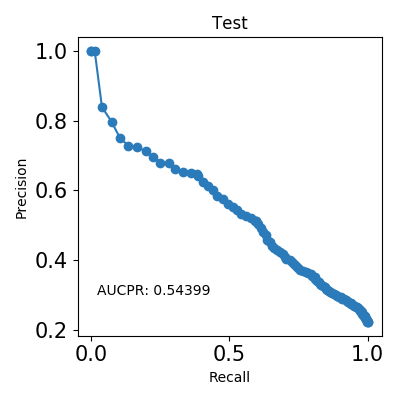
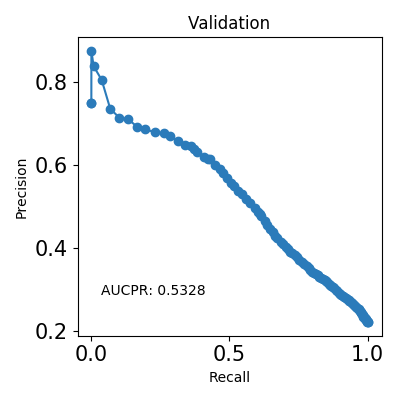
**Test Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 3,922 | 751 | 16% |
| Actual: 1 | 539 | 788 | 41% |

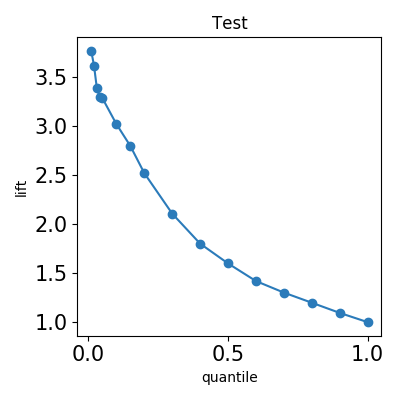
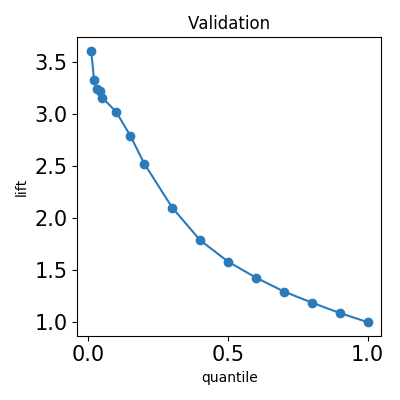
*Receiving Operator Curve*



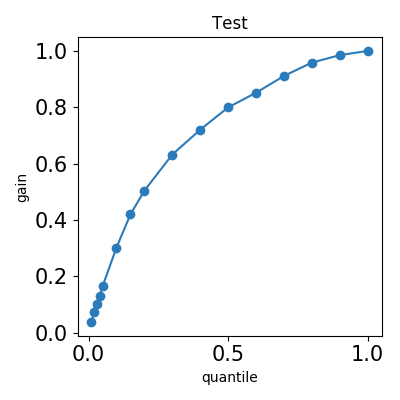
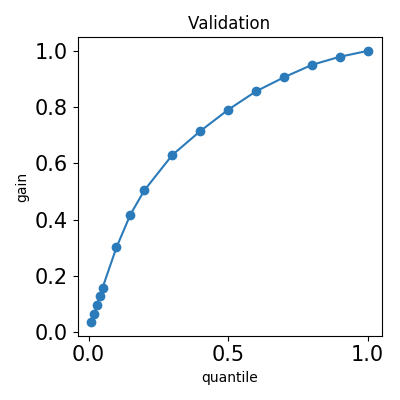
*Precision Recall Curve*



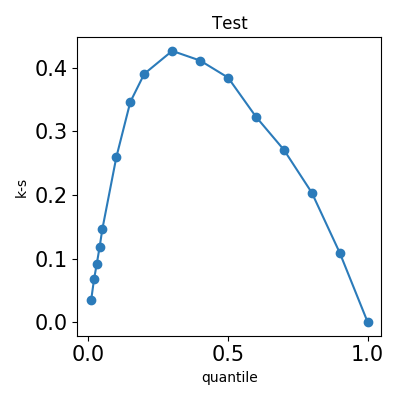
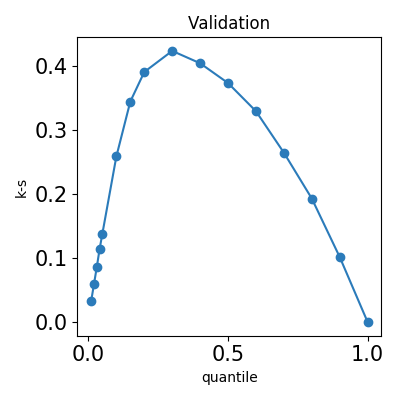
*Cumulative Lift*



*Cumulative Gains*



*Kolmogorov–Smirnov*



## Alternative Models

During the experiment, Driverless AI trained 22 alternative models. The following algorithms were evaluated during the Driverless AI experiment:

|  |  |  |  |
| --- | --- | --- | --- |
| **algorithm** | **package** | **version** | **documentation** |
| gblinear | xgboost | 0.90 | XGBoost: eXtreme Gradient Boosting library. Contributors: https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md |
| gbtree | 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 |
| gbtree | not selected due to low performance during feature evolution stage |
| lightgbm | selected for final model |

## Deployment

For this experiment, both Python and MOJO Scoring Pipelines are available for productionizing the final model pipeline for a given row of data or table of data.

### 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\_854f239a-d5d9-11e9-b3a7-0242ac110002/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, 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).

### MOJO Scoring Pipeline

Note: The MOJO Scoring Pipeline is currently in a beta state. Updates and improvements will continue to be made in subsequent Driverless AI releases. The MOJO Scoring Pipeline is located here:

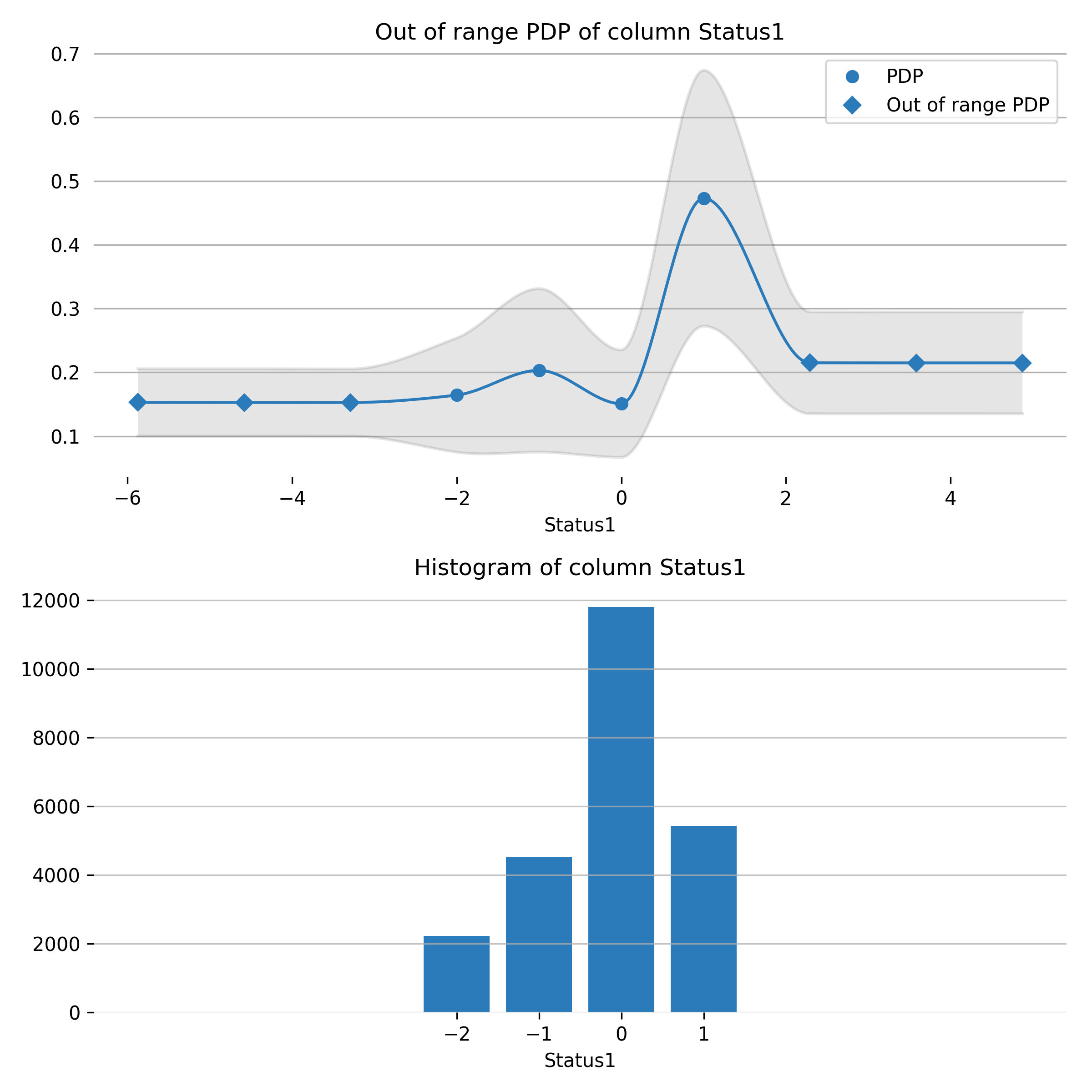
* **h2oai\_experiment\_854f239a-d5d9-11e9-b3a7-0242ac110002/mojo\_pipeline/mojo.zip**

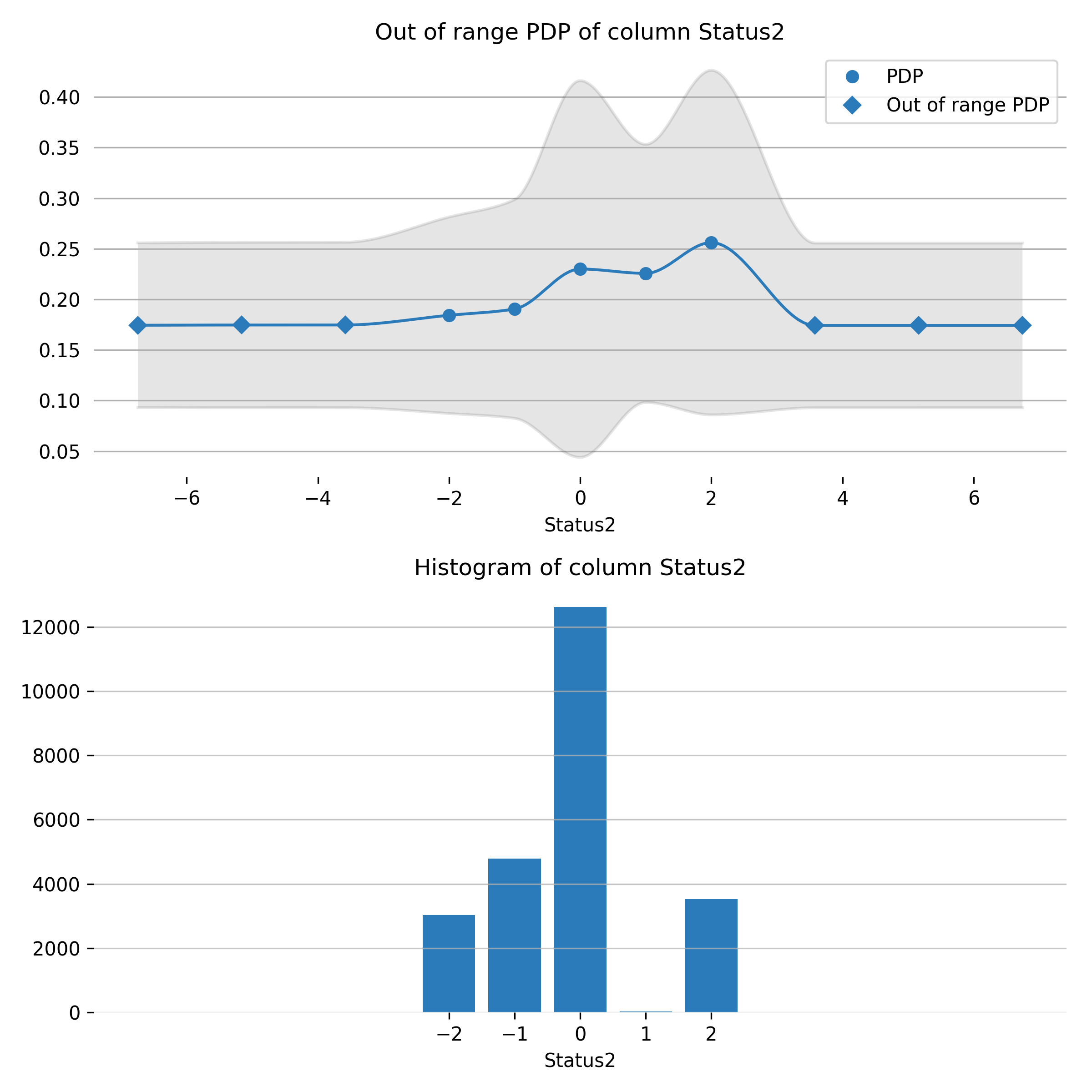
For completed experiments, Driverless AI converts models to MOJOs (Model Objects, Optimized). A MOJO is a scoring engine that can be deployed in any Java environment for scoring in real time.

## Partial Dependence Plots

Partial dependence plots show the partial dependence as a function of specific values for a feature subset. The plots show how machine-learned response functions change based on the values of an input feature of interest, while taking nonlinearity into consideration and averaging out the effects of all other input features. Partial dependence plots enable increased transparency in a model and enable the ability to validate and debug a model by comparing a feature's average predictions across its domain to known standards and reasonable expectations. In the Driverless AI PDP, the y-axis represents the mean response, and a shaded region (for numeric features) or shaded bar (for categorical features) represents 1 standard deviation. Diamond points represent values outside feature intervals seen in data, unseen categorical values or missing values.

The partial dependence plots are shown for the top 2 original variables. The top 2 original variables are chosen based on their Component Based Variable Importance. Partial Dependence computation reached maximum allowed time 20 seconds.

Feature **Status1**

Feature **Status2**

## Appendix

### Final Model Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | LightGBMModel | 1.0 | 4 | 63 | str |

**Model Index: 0**

|  |  |
| --- | --- |
| **parameter** | **value** |
| reg\_lambda | 1.0 |
| learning\_rate | 0.05 |
| gamma | 0 |
| early\_stopping\_rounds | 50 |
| booster | lightgbm |
| gpu\_id | 0 |
| grow\_policy | depthwise |
| min\_child\_samples | 1 |
| boosting\_type | gbdt |
| min\_data\_in\_bin | 1 |
| max\_bin | 256 |
| labels | [0, 1] |
| max\_leaves | 64 |
| random\_state | 98765 |
| max\_depth | 6 |
| reg\_alpha | 0.0 |
| n\_jobs | 2 |
| model\_origin | SEQUENCE |
| min\_child\_weight | 1 |
| num\_class | 1 |
| num\_classes | 2 |
| colsample\_bytree | 0.8 |
| model\_id | 0 |
| monotonicity\_constraints | False |
| silent | True |
| objective | binary:logistic |
| subsample | 0.7 |
| n\_gpus | 1 |
| score\_f\_name | AUC |
| eval\_metric | auc |
| scale\_pos\_weight | 1 |
| n\_estimators | 1200 |
| tree\_method | gpu\_hist |
| max\_delta\_step | 0 |
| nfolds | 4 |