**Driverless AI Experiment: *virahine***

Generated by: megan

Generated on: Thu Apr 18 16:45:39 2019

[Experiment Overview 1](#_Toc532810621)

[Data Overview 3](#_Toc532810622)

[Methodology 4](#_Toc532810623)

[Validation Strategy 9](#_Toc532810624)

[Model Tuning 11](#_Toc532810625)

[Feature Evolution 12](#_Toc532810626)

[Feature Transformations 12](#_Toc532810627)

[Final Model 13](#_Toc532810628)

[Alternative Models 14](#_Toc532810629)

[Deployment 15](#_Toc532810630)

[Appendix 16](#_Toc532810631)

## Experiment Overview

Driverless AI built a stacked ensemble of 1 LightGBMModel, 1 XGBoostModel to predict *default payment next month* given 23 original features from the input dataset *CreditCard-train.csv*. This classification experiment completed in 5 minutes and 36 seconds (0:05:36), using 22 of the 23 original features, and 157 of the 528 engineered features.

### Performance

|  |  |
| --- | --- |
| **Dataset** | **AUC** |
| Internal Validation | 0.782 |
| Test Data | 0.8 |

### Driverless Settings

|  |  |  |  |
| --- | --- | --- | --- |
| Dial Settings | Description | Setting Value | Range of Possible Values |
| Accuracy | Controls accuracy needs of the model | 6 | 1-10 |
| Time | Controls duration of the experiment | 4 | 1-10 |
| Interpretability | Controls complexity of the model | 6 | 1-10 |

### System Specifications

|  |  |  |  |
| --- | --- | --- | --- |
| **System** | **System Memory** | **CPUs** | **GPUs** |
| Docker/Linux | 240 | 32 | 4 |

### Versions

|  |
| --- |
| Driverless AI Version |
| 1.6.1 |

## Data Overview

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

|  |  |  |  |
| --- | --- | --- | --- |
| **data** | **file path** | **number of rows** | **number of columns** |
| training | /data/Kaggle/CreditCard/CreditCard-train.csv | 23,999 | 25 |
| validation | Not provided | None | None |
| testing | /data/Kaggle/CreditCard/CreditCard-test.csv | 6,000 | 25 |

### Training Data

The training data consists of only numeric columns

The summary of the columns is shown below:

#### Numeric Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **data\_type** | **min** | **mean** | **max** | **std** | **unique** | **freq of mode** |
| ID | int | 1.000 | 12,000.000 | 23,999.000 | 6,928.059 | 23,999 | 1 |
| LIMIT\_BAL | int | 10,000.000 | 165,498.716 | 1,000,000.000 | 129,130.743 | 79 | 2,740 |
| SEX | int | 1.000 | 1.628 | 2.000 | 0.483 | 2 | 15,078 |
| EDUCATION | int | 0.000 | 1.847 | 6.000 | 0.780 | 7 | 11,360 |
| MARRIAGE | int | 0.000 | 1.557 | 3.000 | 0.522 | 4 | 12,876 |
| AGE | int | 21.000 | 35.381 | 79.000 | 9.271 | 55 | 1,284 |
| PAY\_0 | int | -2.000 | -0.003 | 8.000 | 1.123 | 11 | 11,738 |
| PAY\_2 | int | -2.000 | -0.123 | 8.000 | 1.201 | 11 | 12,543 |
| PAY\_3 | int | -2.000 | -0.155 | 8.000 | 1.204 | 11 | 12,576 |
| PAY\_4 | int | -2.000 | -0.212 | 8.000 | 1.167 | 11 | 13,250 |
| PAY\_5 | int | -2.000 | -0.253 | 8.000 | 1.137 | 10 | 13,520 |
| PAY\_6 | int | -2.000 | -0.278 | 8.000 | 1.158 | 10 | 12,876 |
| BILL\_AMT1 | int | -165,580.000 | 50,598.929 | 964,511.000 | 72,650.198 | 18,717 | 1,607 |
| BILL\_AMT2 | int | -69,777.000 | 48,648.047 | 983,931.000 | 70,365.396 | 18,367 | 2,049 |
| BILL\_AMT3 | int | -157,264.000 | 46,368.904 | 1,664,089.000 | 68,194.720 | 18,131 | 2,325 |
| BILL\_AMT4 | int | -170,000.000 | 42,369.873 | 891,586.000 | 63,071.455 | 17,719 | 2,547 |
| BILL\_AMT5 | int | -81,334.000 | 40,002.333 | 927,171.000 | 60,345.728 | 17,284 | 2,840 |
| BILL\_AMT6 | int | -339,603.000 | 38,565.267 | 961,664.000 | 59,156.501 | 16,906 | 3,258 |
| PAY\_AMT1 | int | 0.000 | 5,543.098 | 505,000.000 | 15,068.863 | 6,918 | 4,270 |
| PAY\_AMT2 | int | 0.000 | 5,815.529 | 1,684,259.000 | 20,797.444 | 6,839 | 4,362 |
| PAY\_AMT3 | int | 0.000 | 4,969.431 | 896,040.000 | 16,095.929 | 6,424 | 4,853 |
| PAY\_AMT4 | int | 0.000 | 4,743.657 | 497,000.000 | 14,883.555 | 6,028 | 5,200 |
| PAY\_AMT5 | int | 0.000 | 4,783.644 | 417,990.000 | 15,270.704 | 5,984 | 5,407 |
| PAY\_AMT6 | int | 0.000 | 5,189.574 | 528,666.000 | 17,630.719 | 5,988 | 5,846 |

#### Boolean Columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **name** | **data\_type** | **min** | **mean** | **max** | **std** | **freq of max value** |
| default payment next month | bool | False | 0.2237 | True | 0.4167 | 5,369 |

### 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.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 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:

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, 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
   * 28 trained and scored to evaluate features and model parameters
4. **Feature Evolution** 
   * found the best representation of the data for the final model training by creating and evaluating **528** features over **11** iterations
   * 24 trained and scored to further evaluate engineered features
5. **Final Model** 
   * the final model is a stacked ensemble of **1 LightGBMModel, 1 XGBoostModel**
   * 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 automatically created)
   * Python Scoring Pipeline: h2oai\_experiment\_virahine/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 | 3.01 | 0 |
| Model and Feature Tuning | 55.27 | 28 |
| Feature Evolution | 31.84 | 24 |
| Final Pipeline Training | 227.42 | 10 |

### Experiment Settings

Below are the settings selected for the experiment by megan:

**Defined Parameters**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| dataset\_key | tenomiba |
| resumed\_model\_key |  |
| target\_col | default payment next month |
| weight\_col |  |
| fold\_col |  |
| orig\_time\_col |  |
| time\_col | [OFF] |
| is\_classification | True |
| cols\_to\_drop | [] |
| validset\_key |  |
| testset\_key | vesivoro |
| enable\_gpus | True |
| seed | False |
| accuracy | 6 |
| time | 4 |
| interpretability | 6 |
| scorer | AUC |
| is\_timeseries | False |

**Config Overrides**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| recipe | auto |
| enable\_xgboost | auto |
| enable\_lightgbm | auto |
| enable\_rf | auto |
| enable\_glm | auto |
| enable\_tensorflow | off |
| enable\_rulefit | off |
| enable\_ftrl | off |
| parameter\_tuning\_num\_models | -1 |
| fixed\_ensemble\_level | -1 |
| check\_distribution\_shift | True |
| drop\_features\_distribution\_shift\_threshold\_auc | 0.6 |
| target\_transformer | auto |
| enable\_target\_encoding | True |
| time\_series\_recipe | True |
| override\_lag\_sizes |  |
| prob\_lag\_non\_targets | 0.1 |
| make\_python\_scoring\_pipeline | True |
| make\_mojo\_scoring\_pipeline | False |
| rulefit\_max\_num\_rules | -1 |
| feature\_brain\_level | 2 |
| quantile\_imbalanced\_sampling | False |
| holiday\_features | True |
| seed | 1234 |
| force\_64bit\_precision | False |
| min\_num\_rows | 100 |
| max\_orig\_cols\_selected | 10000 |
| nfeatures\_max | -1 |
| feature\_evolution\_data\_size | 100000000 |
| feature\_engineering\_effort | 5 |
| max\_feature\_interaction\_depth | 8 |
| max\_relative\_cardinality | 0.95 |
| string\_col\_as\_text\_threshold | 0.3 |
| tensorflow\_max\_epochs | 10 |
| enable\_tensorflow\_textcnn | False |
| enable\_tensorflow\_textbigru | False |
| enable\_tensorflow\_charcnn | False |
| tensorflow\_max\_epochs\_nlp | 2 |
| min\_dai\_iterations | 0 |
| max\_nestimators | 3000 |
| max\_nestimators\_feature\_evolution\_factor | 0.2 |
| max\_learning\_rate | 0.5 |
| max\_cores | -1 |
| num\_gpus\_per\_model | 1 |
| num\_gpus\_per\_experiment | -1 |
| gpu\_id\_start | 0 |
| compute\_correlation | False |
| high\_correlation\_value\_to\_report | 0.95 |
| dump\_modelparams\_every\_scored\_indiv | False |
| dump\_varimp\_every\_scored\_indiv | False |
| detailed\_traces | False |
| config\_overrides |  |

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 | 40 |
| number of models trained per iteration | 8 |
| early stopping rounds | 5 |
| monotonicity constraint | False |
| number of model tuning model combinations | 27 |
| 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.
* **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, linear 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).

## 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** |
| 9 | lightgbm | 23 | 0.7783 | 1.344 |
| 5 | lightgbm | 23 | 0.7783 | 1.5252 |
| 25 | lightgbm | 182 | 0.7728 | 1.9707 |
| 21 | gbtree | 111 | 0.7723 | 1.4774 |
| 1 | lightgbm | 7 | 0.7722 | 1.3199 |
| 26 | gbtree | 183 | 0.7721 | 1.5232 |
| 18 | lightgbm | 122 | 0.772 | 2.3784 |
| 14 | gbtree | 70 | 0.7717 | 1.2888 |
| 11 | lightgbm | 42 | 0.771 | 1.7418 |
| 19 | lightgbm | 87 | 0.771 | 1.2141 |

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

### gbtree tuning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **tree method** | **grow policy** | **max depth** | **max leaves** | **colsample bytree** | **subsample** | **nfeatures** | **scores** | **training times** |
| gpu\_hist | lossguide | 0.0 | 16.0 | 0.45 | 0.8 | 111 | 0.7723 | 1.4774 |
| gpu\_hist | lossguide | 0.0 | 32.0 | 0.3 | 0.5 | 183 | 0.7721 | 1.5232 |
| gpu\_hist | lossguide | 0.0 | 8.0 | 0.45 | 0.6 | 70 | 0.7717 | 1.2888 |
| gpu\_hist | lossguide | 0.0 | 16.0 | 0.6 | 1 | 76 | 0.7703 | 1.2894 |
| gpu\_hist | depthwise | 9.0 | 0.0 | 0.45 | 0.8 | 43 | 0.7701 | 3.8694 |
| gpu\_hist | depthwise | 9.0 | 0.0 | 0.55 | 0.7 | 88 | 0.7698 | 1.8855 |
| gpu\_hist | lossguide | 0.0 | 64.0 | 0.3 | 0.8 | 80 | 0.7751 | 2.4213 |

### gblinear tuning

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **reg alpha** | **reg lambda** | **n lambda** | **min lambda fraction** | **nfeatures** | **scores** | **training times** |
| 0.025 | 0.05 | 20 | 0.0001 | 19 | 0.7162 | 2.4524 |

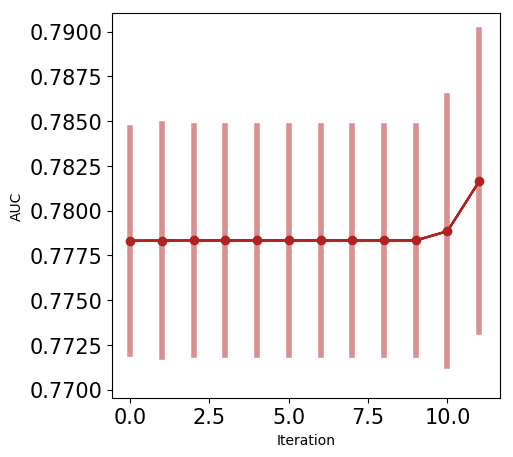
### lightgbm tuning

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **tree method** | **grow policy** | **max depth** | **max leaves** | **colsample bytree** | **subsample** | **nfeatures** | **scores** | **training times** |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.7 | 0.4 | 23 | 0.7783 | 1.344 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.7 | 0.4 | 23 | 0.7783 | 1.5252 |
| gpu\_hist | lossguide | 0.0 | 128.0 | 0.55 | 0.4 | 182 | 0.7728 | 1.9707 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.6 | 0.9 | 7 | 0.7722 | 1.3199 |
| gpu\_hist | depthwise | 4.0 | 0.0 | 0.7 | 0.4 | 122 | 0.772 | 2.3784 |
| gpu\_hist | depthwise | 7.0 | 0.0 | 0.55 | 1 | 42 | 0.771 | 1.7418 |
| gpu\_hist | depthwise | 4.0 | 0.0 | 0.6 | 0.8 | 87 | 0.771 | 1.2141 |
| gpu\_hist | lossguide | 0.0 | 16.0 | 0.5 | 0.8 | 109 | 0.7707 | 1.4807 |
| gpu\_hist | depthwise | 6.0 | 0.0 | 0.7 | 0.4 | 23 | 0.7783 | 1.3487 |
| gpu\_hist | depthwise | 9.0 | 0.0 | 0.2 | 0.5 | 79 | 0.7676 | 1.5332 |

## 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, linear and light gbm models (24) 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:

* **CVTargetEncodeDT**: 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.
* **OHETransformer**: converts a categorical column to a series of boolean features by performing one-hot encoding. The boolean features are used as new features.
* **BulkInteractionsTransformer**: add, divide, multiply, and subtract two numeric columns in the data to create a new feature.
* **ClusterDistTransformer**: clusters selected numeric columns and uses the distance to a specific cluster 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.
* **NumToCatTETransformer**: converts a 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.
* **CVCatNumEncodeDT**: 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** |
| 10\_PAY\_0 | PAY\_0 (original) | None | 1.0 |
| 195\_ClusterTE: ClusterID53: PAY\_0: PAY\_2: PAY\_6.0 | Out-of-fold mean of the response grouped by: ['ClusterID53:PAY\_0:PAY\_2:PAY\_6'] using 5 folds [internal parameters:(10, 5, 10)] (Clustered into 53 clusters) [internal parameters:(53, False, 10, 5, 10)] | Cluster Target Encoding | 0.2663 |
| 208\_NumToCatWoE: LIMIT\_BAL: PAY\_0: PAY\_3.0 | Weight of Evidence for columns ['LIMIT\_BAL', 'PAY\_0', 'PAY\_3'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.2497 |
| 86\_ClusterTE: ClusterID72: LIMIT\_BAL: PAY\_0: PAY\_2.0 | Out-of-fold mean of the response grouped by: ['ClusterID72:LIMIT\_BAL:PAY\_0:PAY\_2'] using 5 folds [internal parameters:(10, 3, 10)] (Clustered into 72 clusters) [internal parameters:(72, False, 10, 3, 10)] | Cluster Target Encoding | 0.2168 |
| 25\_NumToCatWoE: PAY\_3: PAY\_4.0 | Weight of Evidence for columns ['PAY\_3', 'PAY\_4'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.187 |
| 11\_PAY\_2 | PAY\_2 (original) | None | 0.1772 |
| 121\_NumToCatWoE: PAY\_0: PAY\_2.0 | Weight of Evidence for columns ['PAY\_0', 'PAY\_2'] column #0 (numeric columns are bucketed into 100 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.1748 |
| 97\_NumToCatWoE: PAY\_0: PAY\_2: PAY\_3.0 | Weight of Evidence for columns ['PAY\_0', 'PAY\_2', 'PAY\_3'] column #0 (numeric columns are bucketed into 250 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.1663 |
| 1\_BILL\_AMT1 | BILL\_AMT1 (original) | None | 0.1509 |
| 24\_InteractionDiv: LIMIT\_BAL: PAY\_2 | [LIMIT\_BAL] / [PAY\_2] | Interaction | 0.1438 |
| 16\_PAY\_AMT1 | PAY\_AMT1 (original) | None | 0.141 |
| 17\_PAY\_AMT2 | PAY\_AMT2 (original) | None | 0.1332 |
| 189\_ClusterTE: ClusterID96: AGE: BILL\_AMT1: PAY\_0: PAY\_2: PAY\_3: PAY\_AMT4.0 | Out-of-fold mean of the response grouped by: ['ClusterID96:AGE:BILL\_AMT1:PAY\_0:PAY\_2:PAY\_3:PAY\_AMT4'] using 5 folds [internal parameters:(10, 10, 20)] (Clustered into 96 clusters) [internal parameters:(96, False, 10, 10, 20)] | Cluster Target Encoding | 0.1273 |
| 18\_PAY\_AMT3 | PAY\_AMT3 (original) | None | 0.1185 |
| 8\_LIMIT\_BAL | LIMIT\_BAL (original) | None | 0.1145 |
| 0\_AGE | AGE (original) | None | 0.1119 |
| 21\_PAY\_AMT6 | PAY\_AMT6 (original) | None | 0.1089 |
| 19\_PAY\_AMT4 | PAY\_AMT4 (original) | None | 0.1057 |
| 20\_PAY\_AMT5 | PAY\_AMT5 (original) | None | 0.0989 |
| 2\_BILL\_AMT2 | BILL\_AMT2 (original) | None | 0.0963 |
| 3\_BILL\_AMT3 | BILL\_AMT3 (original) | None | 0.0805 |
| 4\_BILL\_AMT4 | BILL\_AMT4 (original) | None | 0.0726 |
| 5\_BILL\_AMT5 | BILL\_AMT5 (original) | None | 0.0712 |
| 194\_NumToCatWoE: LIMIT\_BAL: PAY\_0: PAY\_3.0 | Weight of Evidence for columns ['LIMIT\_BAL', 'PAY\_0', 'PAY\_3'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0647 |
| 14\_PAY\_5 | PAY\_5 (original) | None | 0.0637 |
| 6\_BILL\_AMT6 | BILL\_AMT6 (original) | None | 0.0578 |
| 23\_InteractionSub: BILL\_AMT6: PAY\_2 | [BILL\_AMT6] - [PAY\_2] | Interaction | 0.0519 |
| 143\_ClusterTE: ClusterID68: PAY\_0: PAY\_2.0 | Out-of-fold mean of the response grouped by: ['ClusterID68:PAY\_0:PAY\_2'] using 5 folds [internal parameters:(100, 5, 100)] (Clustered into 68 clusters) [internal parameters:(68, True, 100, 5, 100)] | Cluster Target Encoding | 0.0494 |
| 150\_ClusterTE: ClusterID13: PAY\_0: PAY\_6.0 | Out-of-fold mean of the response grouped by: ['ClusterID13:PAY\_0:PAY\_6'] using 5 folds [internal parameters:(10, 1, None)] (Clustered into 13 clusters) [internal parameters:(13, True, 10, 1, None)] | Cluster Target Encoding | 0.0488 |
| 196\_NumToCatWoE: PAY\_5: PAY\_AMT3.0 | Weight of Evidence for columns ['PAY\_5', 'PAY\_AMT3'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.048 |
| 106\_NumToCatWoE: PAY\_0: PAY\_2.0 | Weight of Evidence for columns ['PAY\_0', 'PAY\_2'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0465 |
| 58\_ClusterTE: ClusterID59: AGE: BILL\_AMT1: PAY\_0: PAY\_2: PAY\_4: PAY\_AMT3.0 | Out-of-fold mean of the response grouped by: ['ClusterID59:AGE:BILL\_AMT1:PAY\_0:PAY\_2:PAY\_4:PAY\_AMT3'] using 5 folds [internal parameters:(100, 1, None)] (Clustered into 59 clusters) [internal parameters:(59, True, 100, 1, None)] | Cluster Target Encoding | 0.0432 |
| 160\_NumToCatWoE: LIMIT\_BAL: PAY\_0: PAY\_2: PAY\_AMT4.0 | Weight of Evidence for columns ['LIMIT\_BAL', 'PAY\_0', 'PAY\_2', 'PAY\_AMT4'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0418 |
| 90\_NumToCatWoE: LIMIT\_BAL: PAY\_0: PAY\_AMT2.0 | Weight of Evidence for columns ['LIMIT\_BAL', 'PAY\_0', 'PAY\_AMT2'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0411 |
| 155\_NumToCatTE: PAY\_0: PAY\_3.0 | Out-of-fold mean of the response grouped by: ['PAY\_0', 'PAY\_3'] using 5 folds [internal parameters:(20, 3, 100)] (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(20, 3, 100)] | Numeric to Categorical Target Encoding | 0.0402 |
| 112\_NumToCatTE: EDUCATION: PAY\_0: PAY\_3.0 | Out-of-fold mean of the response grouped by: ['EDUCATION', 'PAY\_0', 'PAY\_3'] using 5 folds [internal parameters:(20, 1, None)] (numeric columns are bucketed into 100 equally populated bins) [internal parameters:(20, 1, None)] | Numeric to Categorical Target Encoding | 0.0384 |
| 161\_NumToCatWoE: PAY\_2: PAY\_4: PAY\_AMT2.0 | Weight of Evidence for columns ['PAY\_2', 'PAY\_4', 'PAY\_AMT2'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0381 |
| 30\_NumToCatWoE: BILL\_AMT6: PAY\_0: PAY\_2.0 | Weight of Evidence for columns ['BILL\_AMT6', 'PAY\_0', 'PAY\_2'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0346 |
| 171\_NumToCatTE: LIMIT\_BAL: PAY\_2: PAY\_AMT4.0 | Out-of-fold mean of the response grouped by: ['LIMIT\_BAL', 'PAY\_2', 'PAY\_AMT4'] using 5 folds [internal parameters:(10, 3, None)] (numeric columns are bucketed into 10 equally populated bins) [internal parameters:(10, 3, None)] | Numeric to Categorical Target Encoding | 0.0341 |
| 15\_PAY\_6 | PAY\_6 (original) | None | 0.0332 |
| 37\_ClusterTE: ClusterID63: LIMIT\_BAL: PAY\_5: PAY\_AMT2.0 | Out-of-fold mean of the response grouped by: ['ClusterID63:LIMIT\_BAL:PAY\_5:PAY\_AMT2'] using 5 folds [internal parameters:(100, 5, 10)] (Clustered into 63 clusters) [internal parameters:(63, True, 100, 5, 10)] | Cluster Target Encoding | 0.033 |
| 210\_NumToCatTE: PAY\_0: PAY\_6: PAY\_AMT5.0 | Out-of-fold mean of the response grouped by: ['PAY\_0', 'PAY\_6', 'PAY\_AMT5'] using 5 folds [internal parameters:(10, 10, None)] (numeric columns are bucketed into 25 equally populated bins) [internal parameters:(10, 10, None)] | Numeric to Categorical Target Encoding | 0.0304 |
| 75\_ClusterTE: ClusterID59: BILL\_AMT6: LIMIT\_BAL: PAY\_0: PAY\_2: PAY\_5: PAY\_AMT6.0 | Out-of-fold mean of the response grouped by: ['ClusterID59:BILL\_AMT6:LIMIT\_BAL:PAY\_0:PAY\_2:PAY\_5:PAY\_AMT6'] using 5 folds [internal parameters:(100, 5, None)] (Clustered into 59 clusters) [internal parameters:(59, True, 100, 5, None)] | Cluster Target Encoding | 0.0298 |
| 145\_ClusterTE: ClusterID34: PAY\_0: PAY\_2: PAY\_4: PAY\_AMT2.0 | Out-of-fold mean of the response grouped by: ['ClusterID34:PAY\_0:PAY\_2:PAY\_4:PAY\_AMT2'] using 5 folds [internal parameters:(10, 3, 100)] (Clustered into 34 clusters) [internal parameters:(34, True, 10, 3, 100)] | Cluster Target Encoding | 0.0288 |
| 119\_ClusterTE: ClusterID22: BILL\_AMT5: BILL\_AMT6: LIMIT\_BAL: PAY\_0: PAY\_5.0 | Out-of-fold mean of the response grouped by: ['ClusterID22:BILL\_AMT5:BILL\_AMT6:LIMIT\_BAL:PAY\_0:PAY\_5'] using 5 folds [internal parameters:(10, 10, 10)] (Clustered into 22 clusters) [internal parameters:(22, False, 10, 10, 10)] | Cluster Target Encoding | 0.0282 |
| 183\_NumToCatWoE: PAY\_3: PAY\_AMT3: PAY\_AMT5.0 | Weight of Evidence for columns ['PAY\_3', 'PAY\_AMT3', 'PAY\_AMT5'] column #0 (numeric columns are bucketed into 10 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0264 |
| 103\_NumToCatWoE: BILL\_AMT1: BILL\_AMT3: BILL\_AMT6: PAY\_6: PAY\_AMT4: PAY\_AMT6.0 | Weight of Evidence for columns ['BILL\_AMT1', 'BILL\_AMT3', 'BILL\_AMT6', 'PAY\_6', 'PAY\_AMT4', 'PAY\_AMT6'] column #0 (numeric columns are bucketed into 25 equally populated bins) | Numeric to Categorical Weight of Evidence | 0.0256 |
| 197\_NumToCatTE: PAY\_3: PAY\_AMT1.0 | Out-of-fold mean of the response grouped by: ['PAY\_3', 'PAY\_AMT1'] using 5 folds [internal parameters:(20, 5, 100)] (numeric columns are bucketed into 250 equally populated bins) [internal parameters:(20, 5, 100)] | Numeric to Categorical Target Encoding | 0.0246 |
| 12\_PAY\_3 | PAY\_3 (original) | None | 0.0246 |
| 7\_EDUCATION | EDUCATION (original) | None | 0.0243 |

## Final Model

**Pipeline**

Final StackedEnsemble pipeline with ensemble\_level=2 transforming 23 original features -> 179 features in each of 10 models each fit on 5 internal holdout splits then linearly blended

**Details**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | LightGBMModel | 0.8 | 5 | 25 | str |
| 1 | XGBoostModel | 0.2 | 5 | 168 | str |

* Model Index: 0 has a weight of 0.8 in the final ensemble

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **max depth** | **learning rate** | **colsample bytree** | **subsample** | **index** | **tree method** | **grow policy** | **max leaves** |
| LightGBMModel | 6 | 0.04 | 0.7 | 0.4 | 0 | gpu\_hist | depthwise | 64 |

* Model Index: 1 has a weight of 0.2 in the final ensemble

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **max depth** | **learning rate** | **colsample bytree** | **subsample** | **index** | **tree method** | **grow policy** | **max leaves** |
| XGBoostModel | 0 | 0.04 | 0.3 | 0.5 | 1 | gpu\_hist | lossguide | 32 |

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 | 0.82197 | 0.0056423 | 0.8335 | 0.0056423 |
| AUC | \* | higher | 0.78163 | 0.0084403 | 0.80042 | 0.0084403 |
| AUCPR |  | higher | 0.56017 | 0.017099 | 0.57705 | 0.017099 |
| F05 |  | higher | 0.59152 | 0.015634 | 0.59868 | 0.015634 |
| F1 |  | higher | 0.5554 | 0.012527 | 0.55391 | 0.012527 |
| F2 |  | higher | 0.64561 | 0.0095593 | 0.64504 | 0.0095861 |
| GINI |  | higher | 0.56327 | 0.016881 | 0.60085 | 0.016881 |
| LOGLOSS |  | lower | 0.43067 | 0.0089665 | 0.40589 | 0.0089665 |
| MCC |  | higher | 0.43134 | 0.015763 | 0.44565 | 0.015763 |

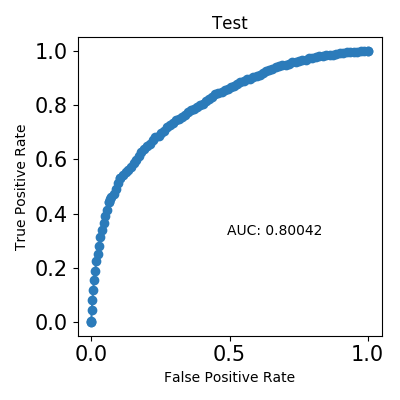
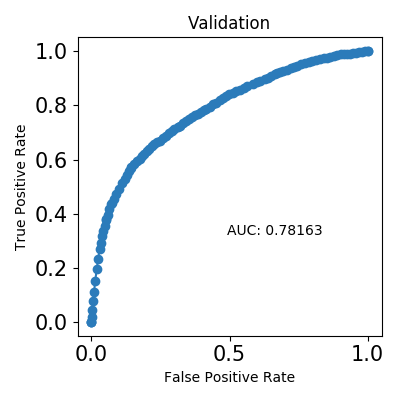
**Validation Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 15,896 | 2,734 | 15% |
| Actual: 1 | 2,293 | 3,076 | 43% |

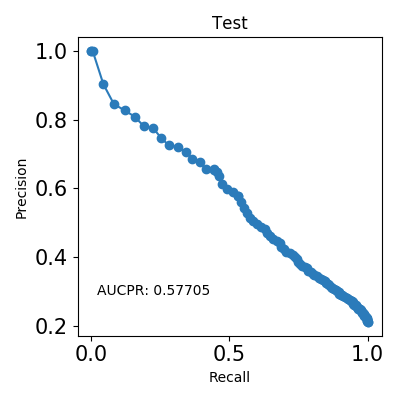
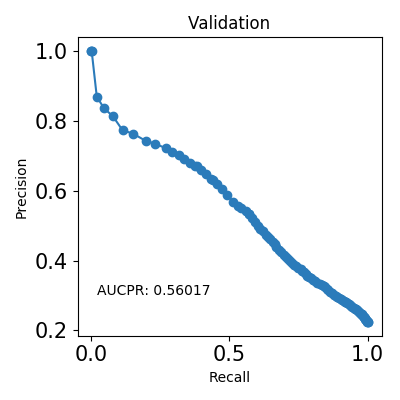
**Test Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 4,122 | 612 | 13% |
| Actual: 1 | 558 | 708 | 44% |

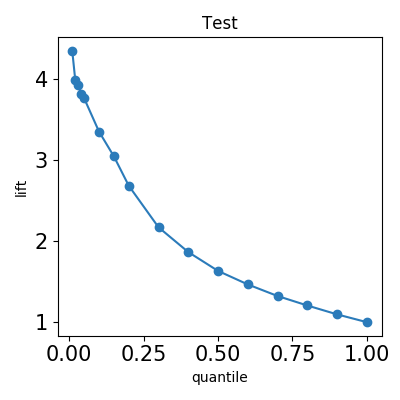
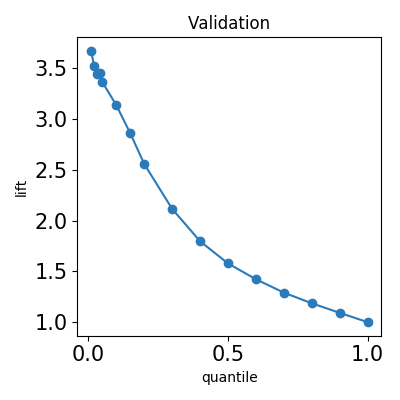
*Receiving Operator Curve*



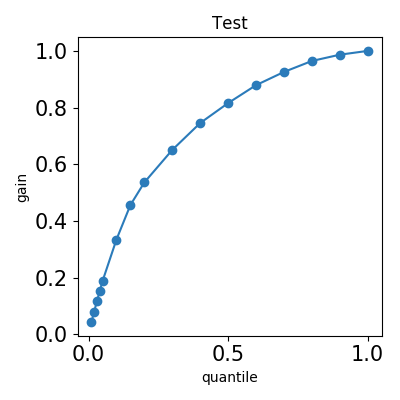
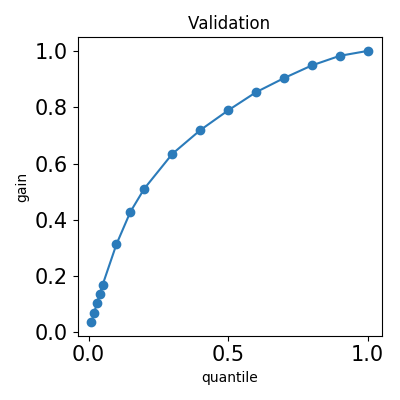
*Precision Recall Curve*



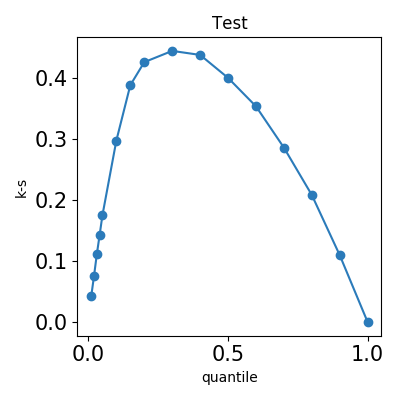
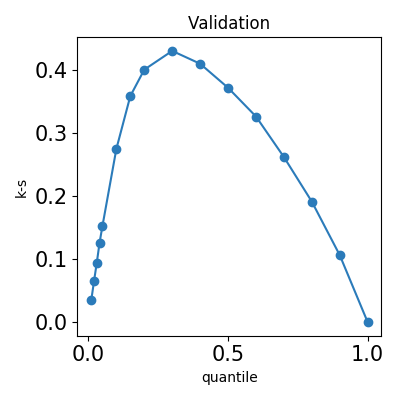
*Cumulative Lift*



*Cumulative Gains*



*Kolmogorov–Smirnov*



## Alternative Models

During the experiment, Driverless AI trained 52 alternative models. The following algorithms were evaluated during the Driverless AI experiment:

|  |  |  |  |
| --- | --- | --- | --- |
| **algorithm** | **package** | **version** | **documentation** |
| gbtree | xgboost | 0.81 | XGBoost: eXtreme Gradient Boosting library. Contributors: https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md |
| gblinear | xgboost | 0.81 | 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 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 |
| gblinear | not selected due to low performance during feature evolution stage |
| gbtree | selected for final model |
| lightgbm | 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\_virahine/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).

## Appendix

### Final Model Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | LightGBMModel | 0.8 | 5 | 25 | str |
| 1 | XGBoostModel | 0.2 | 5 | 168 | str |

**Model Index: 0**

|  |  |
| --- | --- |
| **parameter** | **value** |
| booster | lightgbm |
| min\_child\_samples | 1 |
| scale\_pos\_weight | 1.0 |
| score\_f\_name | AUC |
| early\_stopping\_threshold | 0 |
| subsample | 0.4 |
| reg\_alpha | 0.0 |
| eval\_metric | auc |
| max\_delta\_step | 0.0 |
| gamma | 0.0 |
| num\_classes | 2 |
| max\_drop | 200 |
| reg\_lambda | 10.0 |
| max\_depth | 6 |
| max\_bin | 256 |
| min\_data\_in\_bin | 10 |
| silent | True |
| debug\_verbose | 0 |
| grow\_policy | depthwise |
| early\_stopping\_rounds | 100 |
| skip\_drop | 0.5 |
| random\_state | 147415119 |
| subsample\_freq | 1 |
| n\_gpus | 1 |
| learning\_rate | 0.04 |
| colsample\_bytree | 0.7 |
| rate\_drop | 0.01 |
| gpu\_id | 0 |
| min\_child\_weight | 1 |
| tree\_method | gpu\_hist |
| n\_estimators | 1800 |
| n\_jobs | 1 |
| model\_origin | DefaultIndiv: do\_te:True,interp:11,depth:6,num\_as\_cat:False |
| boosting\_type | gbdt |
| objective | binary |
| monotonicity\_constraints | False |
| model\_id | 0 |
| max\_leaves | 64 |
| nfolds | 5 |

**Model Index: 1**

|  |  |
| --- | --- |
| **parameter** | **value** |
| booster | gbtree |
| min\_child\_samples | 1 |
| scale\_pos\_weight | 1.0 |
| score\_f\_name | AUC |
| early\_stopping\_threshold | 0 |
| subsample | 0.5 |
| reg\_alpha | 0.0 |
| eval\_metric | logloss |
| max\_delta\_step | 3.4699199106 |
| gamma | 0.01 |
| num\_classes | 2 |
| max\_drop | 200 |
| reg\_lambda | 5.0 |
| max\_depth | 0 |
| max\_bin | 256 |
| min\_data\_in\_bin | 1 |
| silent | True |
| debug\_verbose | 0 |
| grow\_policy | lossguide |
| early\_stopping\_rounds | 100 |
| skip\_drop | 0.5 |
| random\_state | 147415119 |
| subsample\_freq | 1 |
| n\_gpus | 1 |
| learning\_rate | 0.04 |
| colsample\_bytree | 0.3 |
| rate\_drop | 0.01 |
| gpu\_id | 0 |
| min\_child\_weight | 1 |
| tree\_method | gpu\_hist |
| n\_estimators | 1800 |
| n\_jobs | 1 |
| model\_origin | RANDOM |
| boosting\_type | gbdt |
| objective | binary:logistic |
| monotonicity\_constraints | False |
| model\_id | 1 |
| max\_leaves | 32 |
| nfolds | 5 |