# Driverless AI Experiment: 3.hatespeech\_vec\_janome

Generated by: yuki.shimada@h2o.ai

Generated on: Thu May 2 12:25:00 2024

[Driverless AI Experiment: 3.hatespeech\_vec\_janome 1](#_Toc55907577)

[Experiment Overview 1](#_Toc55907578)

[Data Overview 1](#_Toc55907579)

[Methodology 1](#_Toc55907580)

[Data Sampling 1](#_Toc55907581)

[Validation Strategy 2](#_Toc55907582)

[Model Tuning 2](#_Toc55907583)

[Feature Evolution 2](#_Toc55907584)

[Feature Transformations 2](#_Toc55907585)

[Final Model 2](#_Toc55907586)

[Alternative Models 2](#_Toc55907587)

[Deployment 2](#_Toc55907588)

[Partial Dependence Plots 2](#_Toc55907589)

[Shapley Values 2](#_Toc55907590)

[Appendix 3](#_Toc55907593)

## Experiment Overview

Driverless AI built a stacked ensemble of 1 LightGBMModel, 1 XGBoostGBMModel to predict label given 6 original features from the input dataset hatespeech\_train\_DAI.csv. This classification experiment completed in 6 minutes and 57 seconds (0:06:57), using 0 of the 6 original features, and 1 of the 1 engineered features.

### Performance

|  |  |
| --- | --- |
| **Dataset** | **F1** |
| Internal Validation | 0.52 |
| Test Data | 0.505 |

### Driverless AI 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 | 3 | 1-10 |
| Interpretability | Controls complexity of the model | 7 | 1-10 |

### System Specifications

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Address** | **System** | **System Memory** | **CPUs** | **GPUs** |
| http://127.0.0.1:12345 | Linux | 248 GB | 32 | 0 |

### Versions

|  |  |
| --- | --- |
|  |  |
| Driverless AI version | 1.10.7 |
| h2o4gpu version | 0.4.2 |
| h2o\_mli version | 1.10.26 |
| mojo2\_runtime version | 2.8.5 |
| daimojo version | 2.8.5 |
| daimojo version\_ppc64le | 2.8.1 |
| procsy version | 0.9.0 |
| pydatatable version | 1.1.0a2217 |
| vis\_data\_server version | 2.1.4 |

## Data Overview

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **data** | **file path** | **file size** | **number of rows** | **number of columns** |
| training | /dai-data/yuki.shimada@h2o.ai/d008d1c0-084b-11ef-aabe-22b0f24f7d4a/hatespeech\_train\_DAI.csv.1714630717.8798265.bin | 2.3 MiB | 4,656 | 7 |
| validation | Not provided | None | None | None |
| testing | /dai-data/yuki.shimada@h2o.ai/c8cbd89e-084b-11ef-aabe-22b0f24f7d4a/hatespeech\_test\_DAI.csv.1714630705.7320695.bin | 307.0 KiB | 600 | 6 |

### Training Data

The training data consists of both numeric and categorical columns.

The summary of the columns is shown below:

#### Numeric Columns

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **logical\_type** | **storage\_type** | **min** | **mean** | **max** | **std** | **unique** | **freq of mode** |
| hold | N/A | int | 0.000 | 2.008 | 4.000 | 1.406 | 5 | 967 |

#### Boolean Columns

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **logical\_type** | **storage\_type** | **min** | **mean** | **max** | **std** | **freq of max value** |
| label | N/A | bool | False | 0.0582 | True | 0.2342 | 271 |

#### Categorical Columns

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **name** | **logical\_type** | **storage\_type** | **unique** | **top** | **freq of top value** |
| id | N/A | str | 4,656 | 001bb4f05 | 1 |
| source | N/A | str | 3 | newsplus | 1,749 |
| text | N/A | str | 4,656 | !aiはぇ～キャベツが一番頑張ってるんか、1位じゃない癖にイキってるホモサピエンスｗｗｗｗｗ | 1 |
| text\_token\_janome | catlabel, text | str | 4,656 | ! ai は ぇ ～ キャベツ が 一番 頑張っ てる ん か 、 1 位 じゃ ない 癖 に イキ | 1 |
| text\_token\_tohokuBertBase | N/A | str | 4,656 | ! a ##i は ぇ ~ キャ ##ベツ が 一番 頑 ##張っ てる ん か 、 1 位 じゃ | 1 |

#### 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 distribution and found the following significant differences:

* Significant difference detected between training and test data distribution for feature <<<text>>> (AUC: 0.995).
* Significant difference detected between training and test data distribution for feature <<<id>>> (AUC: 0.995).
* Significant difference detected between training and test data distribution for feature <<<text\_token\_tohokuBertBase>>> (AUC: 0.995).

## Methodology

This section describes the experiment methodology.

### Assumptions and Limitations

Driverless AI trains all models based on the training data provided (in this case: *hatespeech\_train\_DAI.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 and found significant differences described below:

|  |  |  |
| --- | --- | --- |
| **shift\_col** | **shift\_data\_first** | **shift\_data\_second** |
| text | training | test |
| id | training | test |
| text\_token\_tohokuBertBase | training | test |

### 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 numerical values

* **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, constant and lightgbm models by training models with different parameters

the best parameters are those that generate the largest **F1** on the internal validation data

97 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** features over **12** iterations

trained and scored 96 models to further evaluate engineered features

* **Final Model**
* - the final model is a stacked ensemble of \*\*1 LightGBMModel, 1 XGBoostGBMModel\*\*  
   - the features of these models 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: yuki.shimada@h2o.ai/h2oai\_experiment\_bebee28a-087b-11ef-9f6a-6212b70f1d03/mojo\_pipeline/mojo.zip

Python Scoring Pipeline: yuki.shimada@h2o.ai/h2oai\_experiment\_bebee28a-087b-11ef-9f6a-6212b70f1d03/scoring\_pipeline/scorer.zip

**Models for Optimization**

Driverless AI trained models throughout the experiment 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 | 14.11 | 0 |
| Model and Feature Tuning | 88.58 | 97 |
| Feature Evolution | 54.77 | 96 |
| Final Pipeline Training | 120.33 | 60 |

### Experiment Settings

Below are the settings selected for the experiment by yuki.shimada@h2o.ai. The Defined Parameters represent the high-level parameters.

**Defined Parameters**

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| is\_classification | True |
| enable\_gpus | False |
| seed | False |
| accuracy | 6 |
| time | 3 |
| interpretability | 7 |
| num\_prediction\_periods | None |
| num\_gap\_periods | None |
| is\_timeseries | False |
| is\_image | 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 | 30 |
| number of models trained per iteration | 4 |
| early stopping rounds | 5 |
| monotonicity constraint | True |
| number of model tuning model combinations | 8 |
| number of base learners in ensemble | 12 |
| 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.

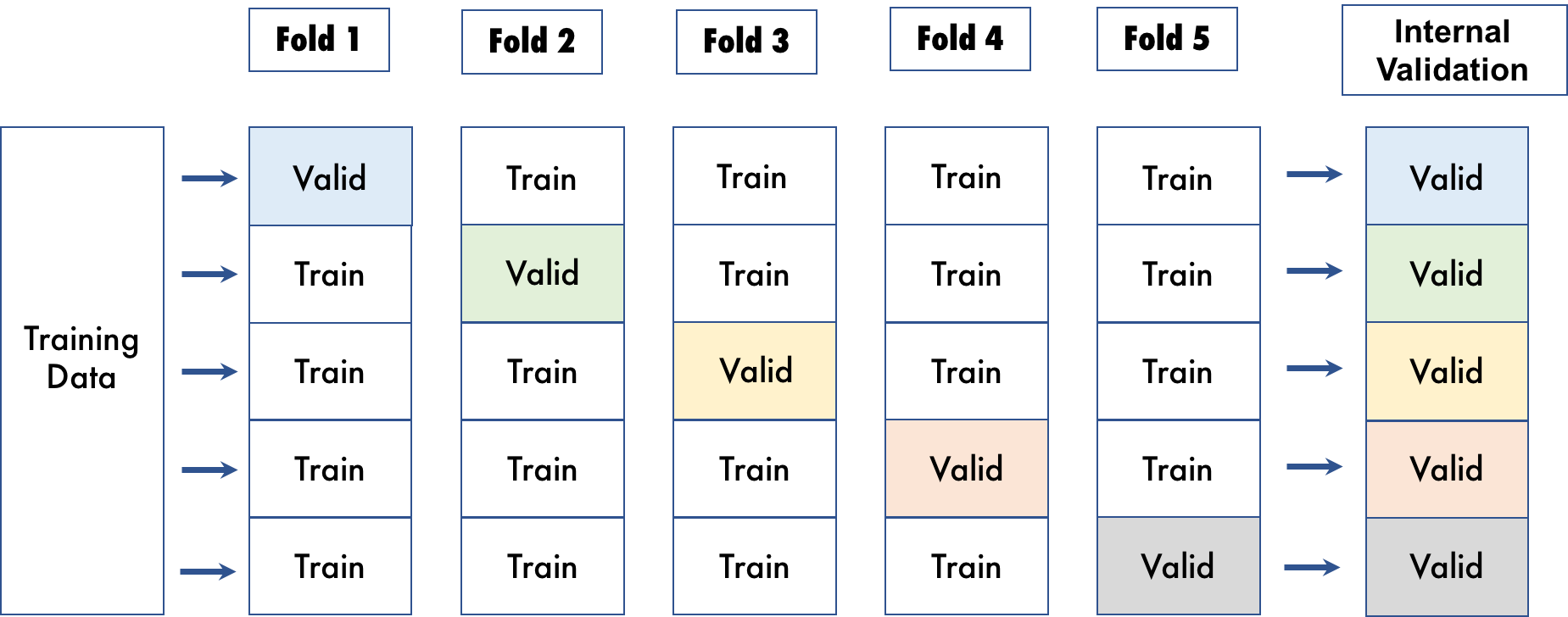
* **number of feature engineering iterations**: the number of iterations performed of feature engineering.
* **number of models evaluated per iteration**: for each feature engineering iteration, Driverless AI trains multiple models, if applicable. Each model is trained with a different set of predictors or features. The goal of this step is to determine which types of features lead to the largest F1 .
* **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. This is off by default given no target variable involved.
* **number of model tuning combinations**: the number of model tuning combinations evaluated to determine the optimal model settings for the xgboost, constant and lightgbm 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). This is off by default for anomaly detection.

## Data Sampling

In Driverless AI, data sampling is a pre-processing step that is done before model training begins; it is not related to sampling done during model training. Driverless AI does not perform data sampling unless the dataset is big or highly imbalanced. Whether a dataset is considered big depends on the experiment’s accuracy setting and the statistical\_threshold\_data\_size\_large config.toml parameter.

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 stages. For the experiment, Driverless AI split the data into 3-fold cross validation, using stratified sampling. With cross validation, the whole dataset is utilized by training 3 models where each model is trained on a different subset of the training data. The visualization below shows how cross validation is utilized to get predictions on hold out data. The visualization shows an example of cross validation with 5 folds. For this experiment, however, 3 folds were created. Note: The cross-validation process was repeated 4 times to ensure the validation metrics are robust since the training data was small. 

## Model Tuning

The table below shows the score and training time of the xgboost, constant and lightgbm models evaluated by Driverless AI. The table shows the parameter tuning models evaluated, ordered based on a combination of largest score and lowest training time.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **job order** | **booster** | **nfeatures** | **training times** | **scores (F1)** |
| 2 | lightgbm | 1 | 0.6782 | 0.4998 |
| 1 | gbtree | 1 | 2.8475 | 0.4994 |
| 3 | gbtree | 1 | 3.0076 | 0.491 |
| 4 | lightgbm | 101 | 4.6461 | 0.4826 |
| 5 | gbtree | 201 | 18.2565 | 0.4892 |
| 6 | lightgbm | 201 | 11.8426 | 0.4854 |
| 0 | lightgbm | 50 | 4.0372 | 0.3825 |
| 7 | constant | 1 | 0.6177 | 0.11 |

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** | **training times** | **scores (F1)** |
| hist | depthwise | 8.0 | 256.0 | 0.8 | 0.7 | 1 | 2.8475 | 0.4994 |
| hist | lossguide | 5.0 | 32.0 | 0.4 | 1.0 | 1 | 3.0076 | 0.491 |
| hist | depthwise | 8.0 | 256.0 | 0.8 | 0.7 | 201 | 18.2565 | 0.4892 |

### constant tuning

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **job order** | **booster** | **nfeatures** | **training times** | **scores (F1)** |
| 7 | constant | 1 | 0.6177 | 0.11 |

### lightgbm tuning

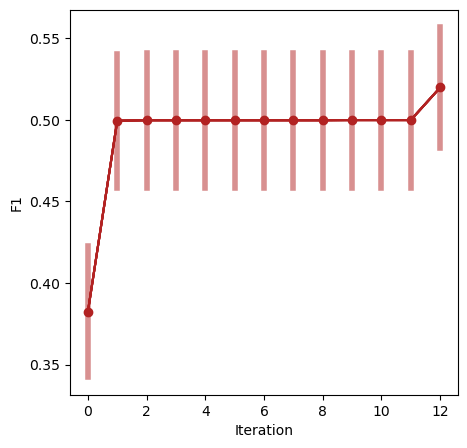
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **tree method** | **grow policy** | **max depth** | **max leaves** | **colsample bytree** | **subsample** | **nfeatures** | **training times** | **scores (F1)** |
|  | depthwise | 8.0 | 256.0 | 0.8 | 0.7 | 1 | 0.6782 | 0.4998 |
|  | depthwise | 7.0 | 128.0 | 1.0 | 0.7 | 101 | 4.6461 | 0.4826 |
|  | depthwise | 8.0 | 256.0 | 0.8 | 0.7 | 201 | 11.8426 | 0.4854 |
|  | depthwise | 8.0 | 256.0 | 0.8 | 0.7 | 50 | 4.0372 | 0.3825 |

## 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, constant and lightgbm models (96) 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:

* **BinnerTransformer**: None
* **ClusterDistTransformer**: the Cluster Distance Transformer clusters selected numeric columns and uses the distance to a specific cluster as a new feature.
* **ClusterTETransformer**: the Cluster Target Encoding Transformer 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**: the Interactions Transformer adds, divides, multiplies, and subtracts two numeric columns in the data to create a new feature.
* **NumToCatTETransformer**: the Numeric to Categorical Target Encoding Transformer 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**: the Numeric to Categorical Weight of Evidence Transformer 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**: the Truncated SVD Transformer trains a Truncated SVD model on selected numeric columns and uses the components of the truncated SVD matrix as new features.
* **CatOriginalTransformer**: the Categorical Original Transformer applies an identity transformation that leaves categorical features as they are. This transformer works with models that can handle non-numeric feature values.
* **CVTargetEncodeTransformer**: the Cross Validation Target Encoding Transformer 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**: the Frequent Transformer calculates the frequency for each value in categorical column(s) and uses this as a new feature. This count can be either the raw count or the normalized count.
* **WeightOfEvidenceTransformer**: the Weight of Evidence Transformer 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**: the Categorical Transformer sorts a categorical column in lexicographical order and uses the order index created as a new feature. This transformer works with models that can handle categorical features.
* **OneHotEncodingTransformer**: the One-hot Encoding transformer converts a categorical column to a series of boolean features by performing one-hot encoding. The boolean features are used as new features.
* **OneHotEncodingUnsupervisedTransformer**: None
* **IsHolidayTransformer**: the Is Holiday Transformer determines if a date column is a holiday. A boolean column indicating if the date is a holiday is added as a new feature. Creates a separate feature for holidays in the United States, United Kingdom, Germany, Mexico, and the European Central Bank. Other countries available in the python Holiday package can be added via the configuration file.
* **DateTimeDiffTransformer**: None
* **DatesTransformer**: the Date Transformer retrieves any date or time values, including: Year, Quarter, Month, Day, Day of Year, Week, Weekday, Hour, Minute, Second.
* **TextTransformer**: the Text Transformer 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.
* **TextLinModelTransformer**: the Text Linear Model Transformer 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.
* **TextOriginalTransformer**: None
* **TimeSeriesTargetEncTransformer**: None
* **ImageVectorizerTransformer**: the Image Vectorizer Transformer converts a column with an image path or URI to the embeddings representation. It utilizes ImageNet pre-trained models with a possible finetuning on the input data.
* **ImageOriginalTransformer**: None
* **CVCatNumEncodeTransformer**: the Cross Validation Categorical to Numeric Encoding Transformer 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**: the Numeric Categorical Target Encoding Transformer 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.

**Dropped Features**

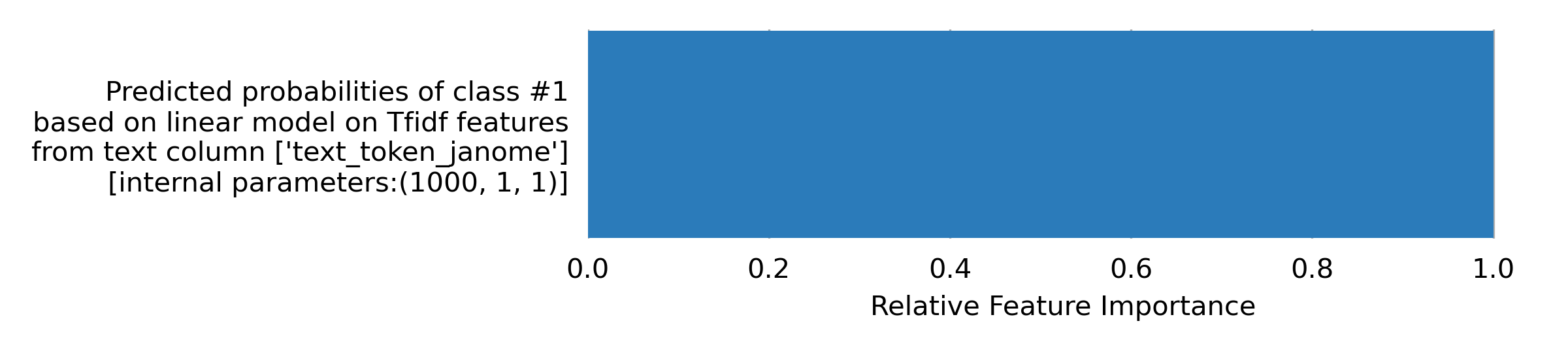
Below is the complete list of dropped features due to numerous reasons.

|  |  |
| --- | --- |
| **Name** | **Reason** |
| hold | User Dropped |
| id | User Dropped |
| source | User Dropped |
| text | User Dropped |
| text\_token\_tohokuBertBase | User Dropped |

## 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 | 0\_TxtTE: text\_token\_janome.0 | Predicted probabilities of class #1 based on linear model on Tfidf features from text column ['text\_token\_janome'] [internal parameters:(1000, 1, 1)] | Bag of Words Regression | 1.0 |



### Advanced NLP and Image Transformations

Driverless AI used advanced NLP/Image transformations in the final model pipeline. Detailed description of each advanced transformer includes the assumptions, limitations and the architecture of the transformer. There may be multiple transformers of the same type in the final pipeline having different architectures.

#### TextLinModelTransformer

Term Frequency - Inverse Document Frequency (TF-IDF) vectorizer creates a representation of texts by constructing a one-hot matrix of word n-grams using the TF-IDF values. These features are used as inputs to get cross-validated predictions from logistic regression model.

**Assumptions**

The order and the position of the words in documents are not relevant for the classification/regression.

**Limitations**

The vocabulary only includes n-grams present in the training set. The algorithm ignores word order. The algorithm cannot take advantage of pretrained word embeddings.

**Sampling**

Tokens with absolute frequency < 3 or relative frequency > 0.9 across the documents are not included in the vocabulary.

**Architecture**

1. TF-IDF parameters:  
 analyzer=word  
 binary=False  
 decode\_error=strict  
 dtype=<class 'numpy.float64'>  
 encoding=utf-8  
 input=content  
 lowercase=True  
 max\_df=0.9  
 max\_features=1000  
 min\_df=3  
 ngram\_range=(1, 1)  
 norm=l2  
 preprocessor=None  
 smooth\_idf=1  
 stop\_words=None  
 strip\_accents=unicode  
 sublinear\_tf=1  
 token\_pattern=(\b\w+\b|[!"#$%&'()\*+,-./:;<=>?@\[\]^\_`{|}~“”¨«»®´·º½¾¿¡§£₤‘’])  
 tokenizer=None  
 use\_idf=1  
 vocabulary=None  
  
2. Logistic regression parameters:  
 C=1,  
 dual=False

**Monotonicity Constraints Details**

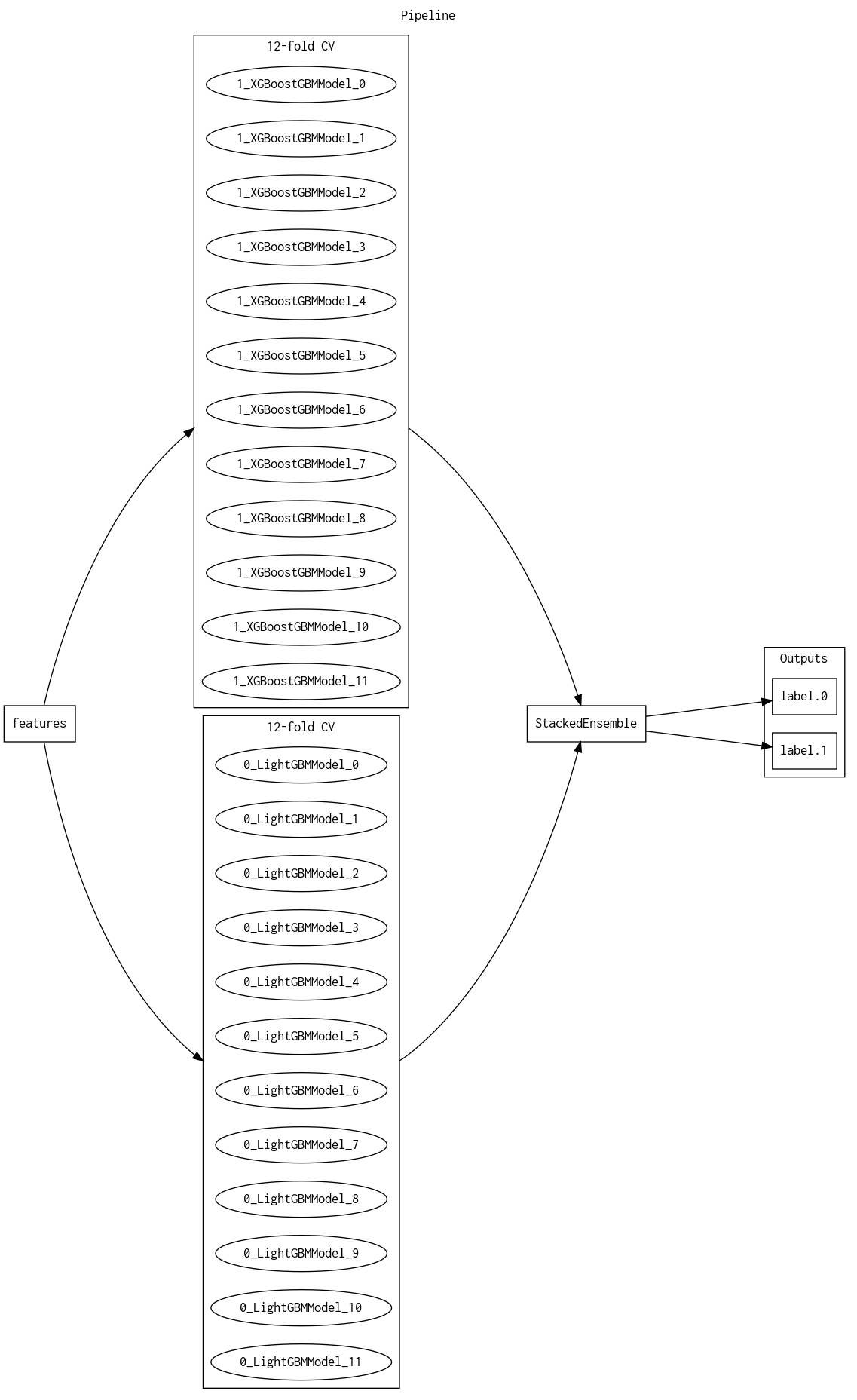
Monotonicity constraints enforce a monotonic relationship between a specified feature and the target prediction. For example, given a model trained to predict housing prices, we might want to enforce that the model predicts higher housing prices with increasing lot size and lower housing prices with increasing neighborhood crime rate.

This experiment enables automatic monotonicity constraints, which mean Driverless AI automatically determines if monotonicity is present and then enforces it through all or part of the modeling pipelines. Depending on the level of feature-target correlation, Driverless AI assigns positive, negative, or no monotonicity constraints. Specifically, monotonicity constraints are enforced if the absolute correlation is greater than 0.1 (the monotonicity correlation threshold set for this experiment). Driverless AI does not enforce monotonicity constraints for features below this correlation threshold.

## Final Model

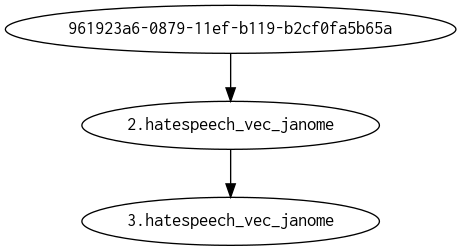
**Pipeline**

Final StackedEnsemble pipeline with ensemble\_level=4 transforming 1 original features -> 1 features in each of 5 models each fit on 12 internal holdout splits then linearly blended:



**Model Lineage**

The following plot shows the experiment lineage for the current experiment 3.hatespeech\_vec\_janome.



**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 | 0.6154 | 12 | 1 | LabelEncoder |
| 1 | XGBoostGBMModel | 0.3846 | 12 | 1 | LabelEncoder |

* Model Index: 0 has a weight of 0.6153845923 in the final ensemble

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **learning rate** | **model class name** | **tree method** | **colsample bytree** | **index** | **max leaves** | **subsample** | **grow policy** | **max depth** |
| LightGBMModel | 0.05 | LightGBMModel |  | 0.8 | 0 | 256 | 0.7 | depthwise | 8 |

* Model Index: 1 has a weight of 0.3846154077 in the final ensemble

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **learning rate** | **model class name** | **tree method** | **colsample bytree** | **index** | **max leaves** | **subsample** | **grow policy** | **max depth** |
| XGBoostGBMModel | 0.05 | XGBoostGBMModel | hist | 0.8 | 1 | 256 | 0.7 | depthwise | 8 |

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

**Performance of Final Model**

Some metrics below require the probabilities to be converted into labels. Predicted labels are generated based on some probability threshold. For example, we may say the label is “YES” if the probability is greater than 0.5.

Choosing this threshold can greatly change the performance metric. This can be seen in the ROC and Precision Recall curves on the Driverless AI UI, where different thresholds on the curve are highlighted, that marks the best Accuracy, the best MCC, and the best F1.

In the Performance Table below, some metrics that require labels have been calculated so that the threshold chosen optimizes the metric. The metrics whose thresholds are optimized are: Accuracy, MCC, F1, F05, and F2. Other metrics that require labels like Precision and Recall choose the threshold that optimizes the F1 because these are trivial to optimize (i.e, Recall will always be perfect when the threshold is 0).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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** |
| F1 | \* | higher | 0.5197738 | 0.03691368 | 0.5048544 | 0.06068729 |
| ACCURACY |  | higher | 0.9504822 | 0.005325954 | 0.9533333 | 0.008537564 |
| AUC |  | higher | 0.9248077 | 0.01241873 | 0.9179267 | 0.02680777 |
| AUCPR |  | higher | 0.4710243 | 0.05034111 | 0.5115523 | 0.08929581 |
| F05 |  | higher | 0.534691 | 0.04952979 | 0.5825243 | 0.08285573 |
| F2 |  | higher | 0.6343166 | 0.03334986 | 0.6355932 | 0.05743045 |
| FDR |  | lower | 0.5483924 | 0.07061602 | 0.6176471 | 0.1628227 |
| FNR |  | lower | 0.3553266 | 0.09259154 | 0.2571429 | 0.1632882 |
| FOR |  | lower | 0.02238652 | 0.005331328 | 0.01691729 | 0.007740038 |
| FPR |  | lower | 0.05202921 | 0.01773502 | 0.07433628 | 0.03268457 |
| GINI |  | higher | 0.8496154 | 0.02483747 | 0.8358534 | 0.05361555 |
| LOGLOSS |  | lower | 0.1363011 | 0.01159121 | 0.1350913 | 0.02033274 |
| MACROAUC |  | higher | 0.9248077 | 0.01241873 | 0.9179267 | 0.02680777 |
| MACROF1 |  | higher | 0.5197738 | 0.03691368 | 0.5048544 | 0.06068729 |
| MACROMCC |  | higher | 0.5037359 | 0.03516783 | 0.4947076 | 0.05943635 |
| MCC |  | higher | 0.5037359 | 0.03516783 | 0.4947076 | 0.05943635 |
| NPV |  | higher | 0.9776135 | 0.005331328 | 0.9830827 | 0.007740038 |
| PRECISION |  | higher | 0.4516076 | 0.07061602 | 0.3823529 | 0.1628227 |
| RECALL |  | higher | 0.6446734 | 0.09259154 | 0.7428571 | 0.1632882 |
| TNR |  | higher | 0.9479708 | 0.01773502 | 0.9256637 | 0.03268457 |

**Validation Confusion Matrix**

Here for the confusion matrix, Driverless AI chose the threshold that maximizes the F1 value on the validation data.

If no validation data was explicitly provided and Driverless AI used cross validation, then the threshold that maximizes the F1 value would be found for each internal fold and, the mean of these thresholds would be calculated.

Note: By default, the confusion matrix that results in best F1 value on some holdout data will be displayed here. This can be configured/changed by assigning different scorer to threshold scorer in recipe settings at https://docs.h2o.ai/driverless-ai/1-10-lts/docs/userguide/expert\_settings/recipes\_settings.html#threshold-scorer.

*Threshold 0.22333048390000002*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 4,146 | 239 | 5% |
| Actual: 1 | 99 | 172 | 37% |

**Test Confusion Matrix**

Here for the confusion matrix, Driverless AI chose the threshold that maximizes the F1 value on the validation data.

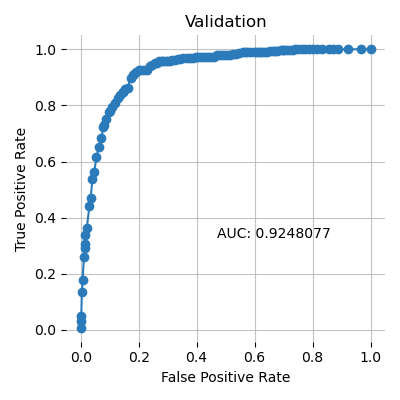
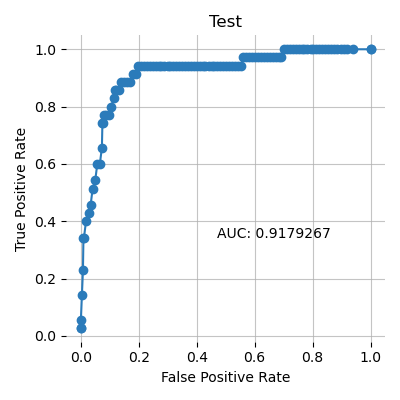
If no validation data was explicitly provided and Driverless AI used cross validation, then the threshold that maximizes the F1 value would be found for each internal fold and, the mean of these thresholds would be calculated.

Note: By default, the confusion matrix that results in best F1 value on some holdout data will be displayed here. This can be configured/changed by assigning different scorer to threshold scorer in recipe settings at https://docs.h2o.ai/driverless-ai/1-10-lts/docs/userguide/expert\_settings/recipes\_settings.html#threshold-scorer.

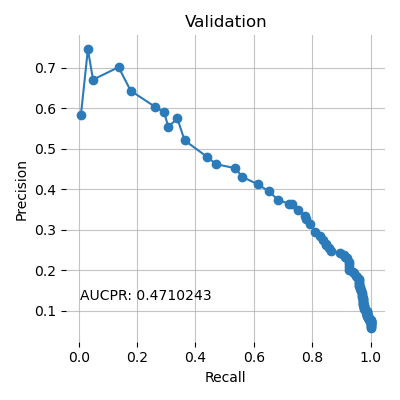
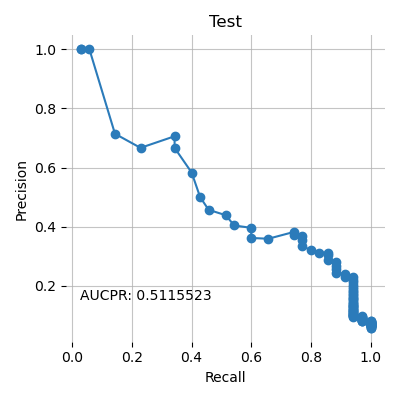
*Threshold 0.22333048390000002*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Predicted: 0** | **Predicted: 1** | **error** |
| Actual: 0 | 533 | 32 | 6% |
| Actual: 1 | 14 | 21 | 40% |

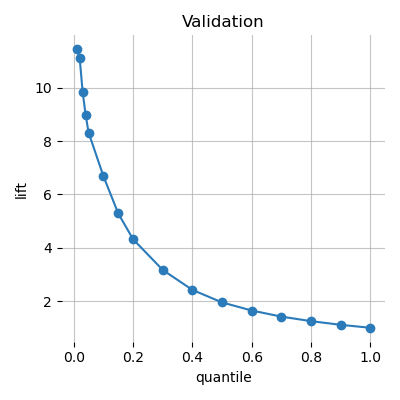
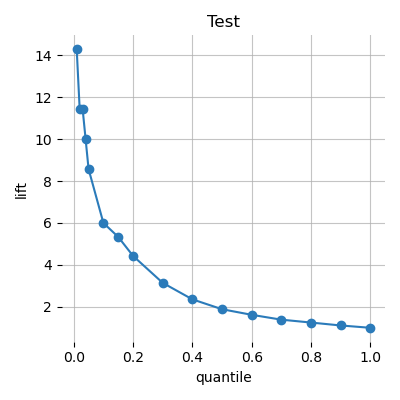
*Receiver Operating Characteristic Curve*

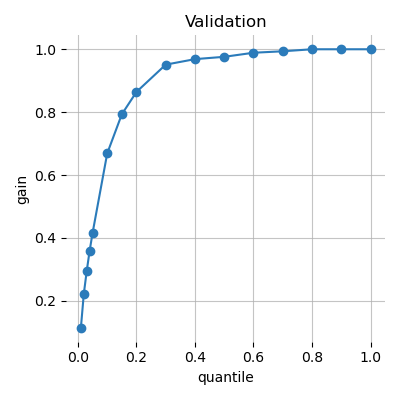
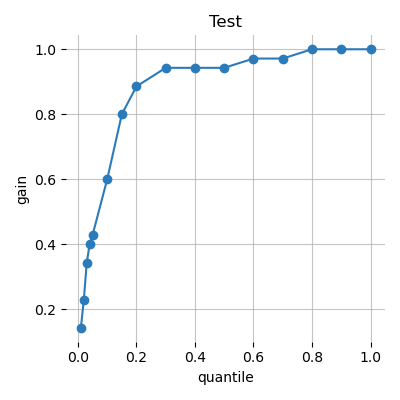
*Precision Recall Curve*

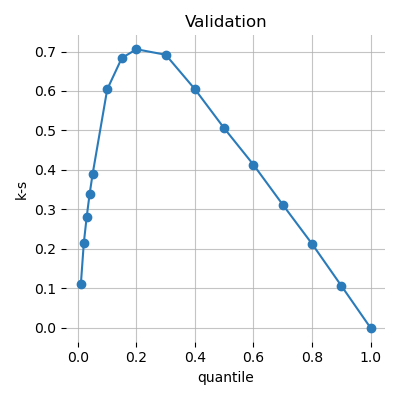
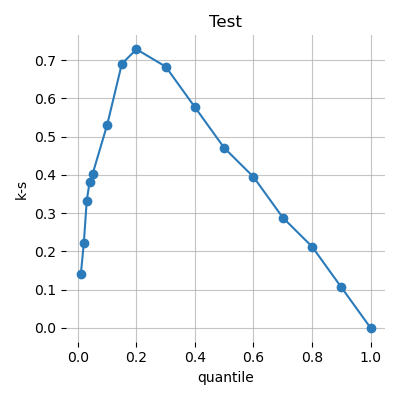
*Cumulative Lift*

*Cumulative Gains*

*Kolmogorov–Smirnov*

## Alternative Models

During the experiment, Driverless AI trained 16 alternative models. The following algorithms were evaluated during the Driverless AI experiment:

|  |  |  |  |
| --- | --- | --- | --- |
| **algorithm** | **package** | **version** | **documentation** |
| gbtree | xgboost | 1.5.0-dev | XGBoost: eXtreme Gradient Boosting library. Contributors: https://github.com/dmlc/xgboost/blob/master/CONTRIBUTORS.md |
| constant | custom package | 1.10.7 | reference model that predicts a constant aimed at minimizing the given scorer |
| lightgbm | lightgbm | 3.3.2.99 | LightGBM, Light Gradient Boosting Machine. Contributors: https://github.com/microsoft/LightGBM/graphs/contributors. |

Driverless AI can evaluate an array of algorithms, including but not limited to 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** |
| gblinear | algorithm not evaluated due to experiment configuration |
| decision tree | algorithm not evaluated due to experiment configuration |
| 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 |
| gbtree | selected for final model |
| 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:

* **yuki.shimada@h2o.ai/h2oai\_experiment\_bebee28a-087b-11ef-9f6a-6212b70f1d03/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).

### 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:

* **yuki.shimada@h2o.ai/h2oai\_experiment\_bebee28a-087b-11ef-9f6a-6212b70f1d03/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.

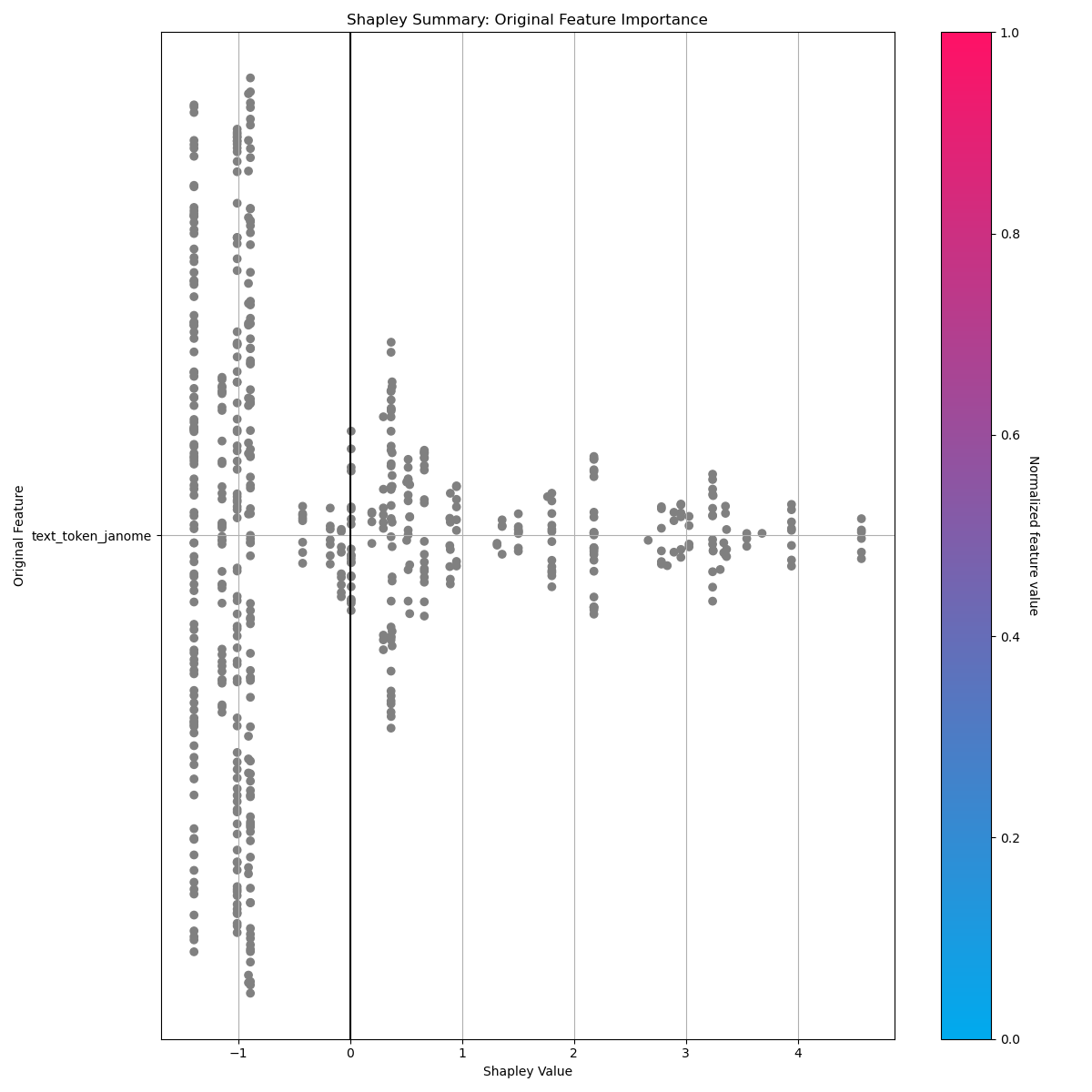
The partial dependence plots are shown for the top 0 original variables. The top 0 original variables are chosen based on their Component Based Variable Importance.

## Shapley Values

#### Shapley Contributions on the Test Dataset

Shapley explanations are a technique with credible theoretical support that presents consistent global and local feature contributions. For regression problems, local Shapley feature contributions plus the bias term sum to the final model’s prediction. For classification problems, they sum to the prediction before applying the link function.

This section uses Driverless AI’s Naive Shapley method to calculate local Shapley explanations for original features. These explanations are approximation for the original features and are based on how often the features are used in transformed features, and how important those transformed features are to the final model. The importance of each transformed feature is distributed equally to all original features that helped create it. This is then summed for each original feature.



## Appendix

### Final Model Details

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Index** | **Type** | **Model Weight** | **Num Folds** | **Fitted features** | **Target Transformer** |
| 0 | LightGBMModel | 0.6154 | 12 | 1 | LabelEncoder |
| 1 | XGBoostGBMModel | 0.3846 | 12 | 1 | LabelEncoder |

**Model Index: 0**

|  |  |
| --- | --- |
| **parameter** | **value** |
| bagging\_seed | 316557807 |
| base\_score |  |
| booster | lightgbm |
| boosting\_type | gbdt |
| categorical\_feature |  |
| class\_weight |  |
| colsample\_bytree | 0.8 |
| debug\_verbose |  |
| deterministic | False |
| device\_type | cpu |
| disable\_gpus | True |
| early\_stopping\_rounds | 150 |
| early\_stopping\_threshold | 0.0 |
| enable\_early\_stopping\_rounds | True |
| eval\_metric | binary |
| feature\_fraction\_seed | 316557806 |
| gamma |  |
| gpu\_id | 0 |
| grow\_policy | depthwise |
| imbalance\_info\_dict | {} |
| importance\_type | gain |
| label\_counts | [4385, 271] |
| labels | [0, 1] |
| learning\_rate | 0.05 |
| max\_bin | 249 |
| max\_delta\_step | 0.0 |
| max\_depth | 8 |
| max\_leaves | 256 |
| min\_child\_samples | 1 |
| min\_child\_weight | 0.001 |
| min\_data\_in\_bin | 1 |
| min\_split\_gain | 0.0 |
| model\_class\_name | LightGBMModel |
| model\_id | 0 |
| monotone\_constraints | [1] |
| monotone\_constraints\_method | intermediate |
| monotone\_penalty | 0.0 |
| monotonicity\_constraints | True |
| n\_estimators | 1000 |
| n\_gpus | 0 |
| n\_jobs | 1 |
| nthread |  |
| num\_class | 1 |
| num\_classes | 2 |
| num\_leaves | 256 |
| num\_threads | 1 |
| objective | binary |
| random\_state | 316557805 |
| reg\_alpha | 0.0 |
| reg\_lambda | 0.0 |
| scale\_pos\_weight | 1.0 |
| score\_f\_name | F1 |
| seed | 316557805 |
| silent |  |
| subsample | 0.7 |
| subsample\_for\_bin | 200000 |
| subsample\_freq | 1 |
| tree\_method |  |
| verbose | -1 |
| nfolds | 12 |

**Model Index: 1**

|  |  |
| --- | --- |
| **parameter** | **value** |
| bagging\_seed |  |
| base\_score | 0.5 |
| booster | gbtree |
| boosting\_type |  |
| categorical\_feature |  |
| class\_weight |  |
| colsample\_bytree | 0.8 |
| debug\_verbose | 0 |
| deterministic |  |
| device\_type |  |
| disable\_gpus | True |
| early\_stopping\_rounds | 150 |
| early\_stopping\_threshold |  |
| enable\_early\_stopping\_rounds | True |
| eval\_metric | logloss |
| feature\_fraction\_seed |  |
| gamma | 0.0 |
| gpu\_id |  |
| grow\_policy | depthwise |
| imbalance\_info\_dict | {} |
| importance\_type | gain |
| label\_counts | [4385, 271] |
| labels | [0, 1] |
| learning\_rate | 0.05 |
| max\_bin | 256 |
| max\_delta\_step | 0.0 |
| max\_depth | 8 |
| max\_leaves | 256 |
| min\_child\_samples |  |
| min\_child\_weight | 1.0 |
| min\_data\_in\_bin |  |
| min\_split\_gain |  |
| model\_class\_name | XGBoostGBMModel |
| model\_id | 1 |
| monotone\_constraints | (1) |
| monotone\_constraints\_method |  |
| monotone\_penalty |  |
| monotonicity\_constraints | True |
| n\_estimators | 1000 |
| n\_gpus | 0 |
| n\_jobs | 1 |
| nthread | 1 |
| num\_class | 1 |
| num\_classes |  |
| num\_leaves |  |
| num\_threads |  |
| objective | binary:logistic |
| random\_state |  |
| reg\_alpha | 0.0 |
| reg\_lambda | 0.0 |
| scale\_pos\_weight | 1.0 |
| score\_f\_name | F1 |
| seed | 316557805 |
| silent | 1 |
| subsample | 0.7 |
| subsample\_for\_bin |  |
| subsample\_freq |  |
| tree\_method | hist |
| verbose |  |
| nfolds | 12 |

### Config Overrides

The Config Overrides represent the fine-control parameters.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| glm\_optimal\_refit | False |
| tournament\_remove\_worse\_than\_constant\_before\_evolution | False |
| max\_runtime\_minutes | 60 |
| tournament\_keep\_absolute\_ok\_scores\_before\_evolution\_model\_factor | 1.1 |
| auth\_oidc\_introspection\_endpoint\_url | https://auth.internal.dedicated.h2o.ai/auth/realms/hac/protocol/openid-connect/token/introspect |
| auth\_oidc\_issuer\_url | https://auth.internal.dedicated.h2o.ai/auth/realms/hac |
| tournament\_remove\_poor\_scores\_before\_evolution\_model\_factor | 1.1 |
| h2o\_storage\_tls\_cert\_path | /pki/client/client-cert |
| base\_url | /workspaces/default/daiEngines/dai-small-7/ |
| virtual\_cores\_per\_physical\_core | 2 |
| ohe\_bin\_list | [10, 50, 100] |
| http\_cookie\_attributes | {'path': '/workspaces/default/daiEngines/dai-small-7'} |
| prob\_lagsinteraction | 0.2 |
| h2o\_drive\_endpoint\_url | http://h2o-drive.drive:8081 |
| h2o\_mlops\_ui\_url | https://mlops.internal.dedicated.h2o.ai |
| auth\_openid\_scope | openid offline\_access profile ai.h2o.storage |
| api\_token\_introspection\_enabled | True |
| prob\_lag\_non\_targets | 0.1 |
| included\_transformers | ['AggregatorTransformer', 'AutovizRecommendationsTransformer', 'BERTTransformer', 'BinnerTransformer', 'CVCatNumEncodeTransformer', 'CVTargetEncodeTransformer', 'CatOriginalTransformer', 'CatTransformer', 'ClusterDistTransformer', 'ClusterIdAllNumTransformer', 'ClusterTETransformer', 'DateOriginalTransformer', 'DateTimeDiffTransformer', 'DateTimeOriginalTransformer', 'DatesTransformer', 'EwmaLagsTransformer', 'FrequentTransformer', 'ImageOriginalTransformer', 'ImageVectorizerTransformer', 'InteractionsTransformer', 'IsHolidayTransformer', 'IsolationForestAnomalyAllNumericTransformer', 'IsolationForestAnomalyNumCatAllColsTransformer', 'IsolationForestAnomalyNumCatTransformer', 'IsolationForestAnomalyNumericTransformer', 'LagsAggregatesTransformer', 'LagsInteractionTransformer', 'LagsTransformer', 'LexiLabelEncoderTransformer', 'MeanTargetTransformer', 'NumCatTETransformer', 'NumToCatTETransformer', 'NumToCatWoEMonotonicTransformer', 'NumToCatWoETransformer', 'OneHotEncodingTransformer', 'OneHotEncodingUnsupervisedTransformer', 'OriginalTransformer', 'RobustScalerTransformer', 'StandardScalerTransformer', 'StringConcatTransformer', 'TextBiGRUTransformer', 'TextCNNTransformer', 'TextCharCNNTransformer', 'TextLinModelTransformer', 'TextOriginalTransformer', 'TextTransformer', 'TimeSeriesTargetEncSimpleTransformer', 'TimeSeriesTargetEncTransformer', 'TruncSVDAllNumTransformer', 'TruncSVDNumTransformer', 'WeightOfEvidenceTransformer'] |
| cv\_in\_cv\_overconfidence\_protection | on |
| last\_recipe | auto |
| auth\_openid\_token\_introspection\_url | http://localhost:9876/introspect |
| auth\_oidc\_post\_logout\_url | https://enginemanager.internal.dedicated.h2o.ai/workspaces/default/daiEngines/dai-small-7/oidc/login |
| auth\_oidc\_username\_claim | preferred\_username |
| prob\_default\_lags | 0.2 |
| auth\_openid\_use\_objectpath\_match | True |
| skip\_record\_h2o4gpu\_version | True |
| application\_id | dai\_7 |
| enable\_startup\_checks | False |
| included\_models | ['Aggregator', 'Constant', 'DecisionTree', 'FTRL', 'GLM', 'ImageAuto', 'ImbalancedLightGBM', 'ImbalancedXGBoostGBM', 'IsolationForestAnomaly', 'KMeans', 'KMeansFreq', 'KMeansOHE', 'LightGBM', 'LightGBMDask', 'RuleFit', 'TensorFlow', 'TextALBERT', 'TextBERT', 'TextCamemBERT', 'TextDistilBERT', 'TextMultilingualBERT', 'TextRoBERTa', 'TextXLM', 'TextXLMRoberta', 'TextXLNET', 'TorchGrowNet', 'TruncSVD', 'Unsupervised', 'UnsupervisedGeneral', 'UnsupervisedOneAll', 'UnsupervisedOneOutput', 'XGBoostDart', 'XGBoostDartDask', 'XGBoostGBM', 'XGBoostGBMDask', 'XGBoostRF', 'XGBoostRFDask', 'ZeroInflatedLightGBM', 'ZeroInflatedXGBoost'] |
| h2o\_storage\_tls\_key\_path | /pki/client/client-key |
| last\_exclusive\_mode | safe |
| override\_virtual\_cores | 8 |
| cuda\_visible\_devices |  |
| h2o\_storage\_address | mlops-storage.mlops:9999 |
| fast\_startup\_callback\_delay | 30.0 |
| h2o\_storage\_projects\_enabled | True |
| auth\_oidc\_identity\_source | id\_token |
| tournament\_remove\_poor\_scores\_before\_final\_model\_factor | 1.1 |
| threshold\_scorer | F1 |
| authentication\_method | oidc |
| enable\_xgboost\_h2o\_recipes | False |
| worker\_ip | 10.3.17.7 |
| auth\_openid\_use\_objectpath\_expression | $.sub is "bc34ef4d-ff29-4375-8bab-04d860fcac32" |
| server\_ip | 10.3.17.7 |
| auth\_openid\_redirect\_uri | https://enginemanager.internal.dedicated.h2o.ai/workspaces/default/daiEngines/dai-small-7/oidc/callback |
| fast\_startup | True |
| num\_gpus\_per\_model | 0 |
| prob\_lagsaggregates | 0.2 |
| check\_system | False |
| num\_gpus\_per\_experiment | 0 |
| disablelogout | True |
| disk\_limit\_gb | 0 |
| enable\_quick\_benchmark | False |
| how\_started | GUI |
| skip\_track\_system\_info\_startup | True |
| auth\_openid\_access\_token\_expiry\_key |  |
| hac\_link\_url | https://internal.dedicated.h2o.ai |
| h2o\_storage\_tls\_ca\_path | /pki/ca/ca-cert |
| fast\_startup\_callback\_delay\_step | 5 |
| auth\_oidc\_token\_endpoint\_url | https://auth.internal.dedicated.h2o.ai/auth/realms/hac/protocol/openid-connect/token |
| included\_scorers | ['ACCURACY', 'AUC', 'AUCPR', 'CALINSKIHARABASZ', 'DAVIESBOULDIN', 'F05', 'F1', 'F2', 'FDR', 'FNR', 'FOR', 'FPR', 'GINI', 'LOGLOSS', 'MACROAUC', 'MACROF1', 'MACROMCC', 'MAE', 'MAPE', 'MCC', 'MER', 'MSE', 'NPV', 'PRECISION', 'R2', 'R2COD', 'RECALL', 'RMSE', 'RMSLE', 'RMSPE', 'SILHOUETTE', 'SMAPE', 'TNR', 'UNSUPERVISED'] |
| memory\_limit\_gb | 0 |