SageMaker Model Quality Report

This report contains model insights and model quality information for candidate **WeightedEnsemble_L2_FULL**. The candidate was generated by the AutoML job **widschallenge**.

The **WeightedEnsemble_L2_FULL** candidate is a trained **binary** model whose objective is to **Maximize** the **"f1"** quality metric.

The F1 score is the harmonic mean of the precision and recall. It is used for binary classification into classes traditionally referred to as positive and negative. Predictions are said to be true when they match their actual (correct) class, and false when they do not.

Precision is the ratio of the true positive predictions to all positive predictions (including the false positives) in a data set, and it measures the quality of the prediction when it predicts the positive class. Recall (or sensitivity) is the ratio of the true positive predictions to all actual positive instances, and it measures how completely a model predicts the actual class members in a data set. The standard F1 score weighs precision and recall equally. Depending on specific aspects of a problem, the standard F1 score determines which metric is paramount. F1 scores vary between zero (0) and one (1): The best possible performance will score 1 , whereas 0 indicates the worst.

Contents

Autopilot job details

Title	Value
Autopilot candidate name	WeightedEnsemble_L2_FULL
Autopilot job name	widschallenge
Problem type	binary
Objective metric	f1
Optimization direction	Maximize

Model quality report

Model quality information is generated by the Autopilot Local Model Insights. This report is for a **binary** problem. **2582** rows were included in the evaluation dataset. The evaluation time occurred at **2024-02-12 06:54:47**.

Metrics table

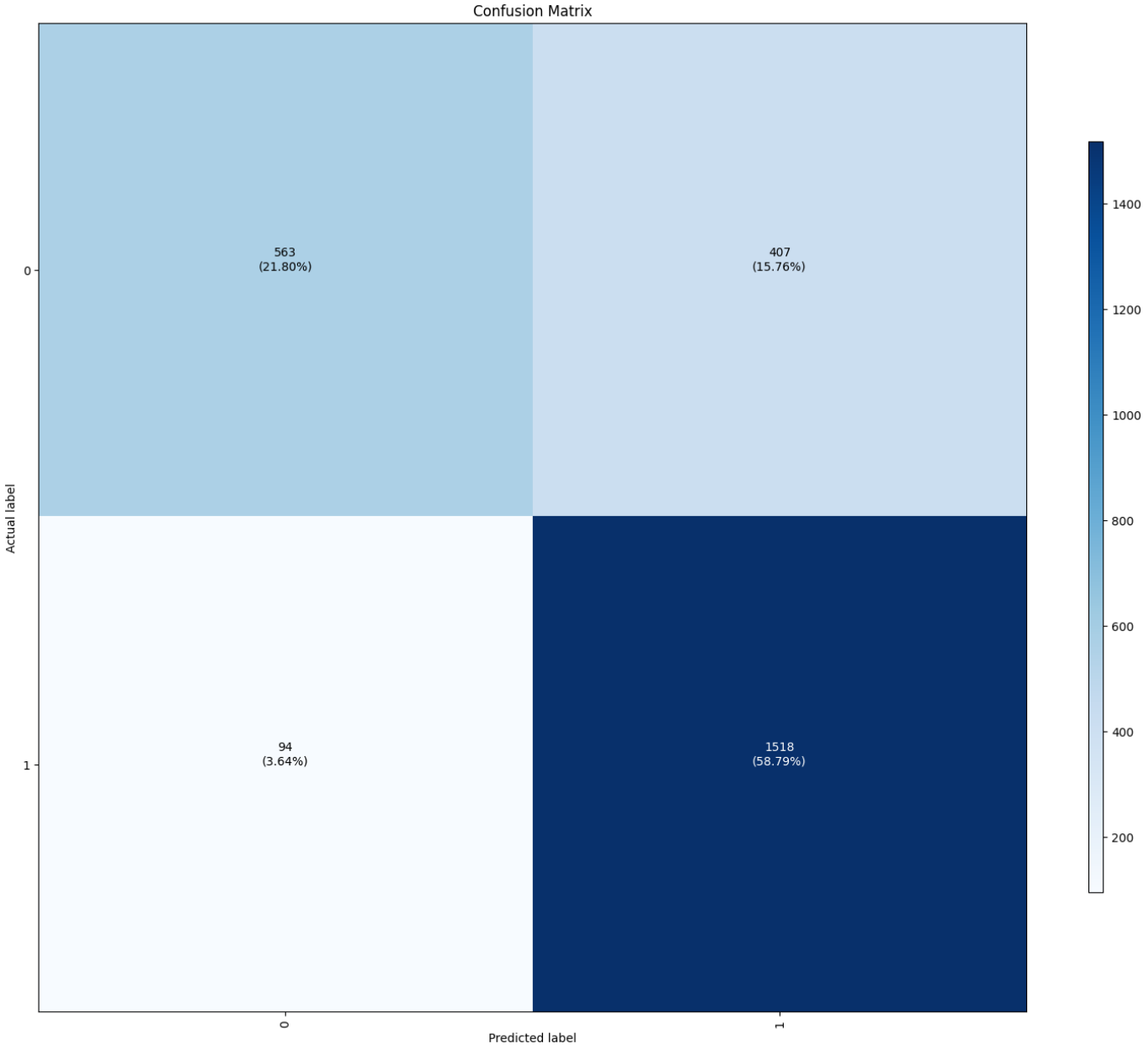
Metric Name	Value	Standard Deviation
recall	0.580412	0.002630
precision	0.856925	0.007659
accuracy	0.805964	0.002521
f0_5	0.782379	0.005120
f1	0.692071	0.003001
f2	0.620454	0.002476
recall_best_constant_classifier	0.000000	0.000000
precision_best_constant_classifier	0.000000	0.000000
accuracy_best_constant_classifier	0.624322	0.004468
f0_5_best_constant_classifier	0.000000	0.000000
f1_best_constant_classifier	0.000000	0.000000
f2_best_constant_classifier	0.000000	0.000000
true_positive_rate	0.580412	0.002630
true_negative_rate	0.941687	0.003415
false_positive_rate	0.058313	0.003415
false_negative_rate	0.419588	0.002630
auc	0.791367	0.003835
au_prc	0.751572	0.005523

Note The values of the performance metrics in this table may differ from the values reported by Autopilot. The differences tend to appear when training on smaller datasets. The values for the metrics in the table use all the training data once to estimate the performance of a model. Autopilot scores are calculated using k-fold cross-validation resampling method that train a machine learning algorithm on different subsets of the dataset. A score is then calculated for overall performance by averaging the resulting performance metrics for each trial.

Confusion matrix

The **confusion matrix** provides a way to visualize the accuracy of the predictions made by a classification model. The confusion matrix is a table that contains the percentages of correct and incorrect predictions for the actual labels. Each row in the confusion matrix indicates how an actual label was classified by the label predicted by the model. The percentage of accurate predictions are on the diagonal, from the upper-left to the lower-right corner. The off-diagonal percentages indicate the types of misclassifications that the model is predicting. These incorrect predictions are the confusion values.

NOTE: If a row shows Nan , it means the validation dataset doesn't have a row for that label.



The area under the receiver operating characteristic curve (AUC)

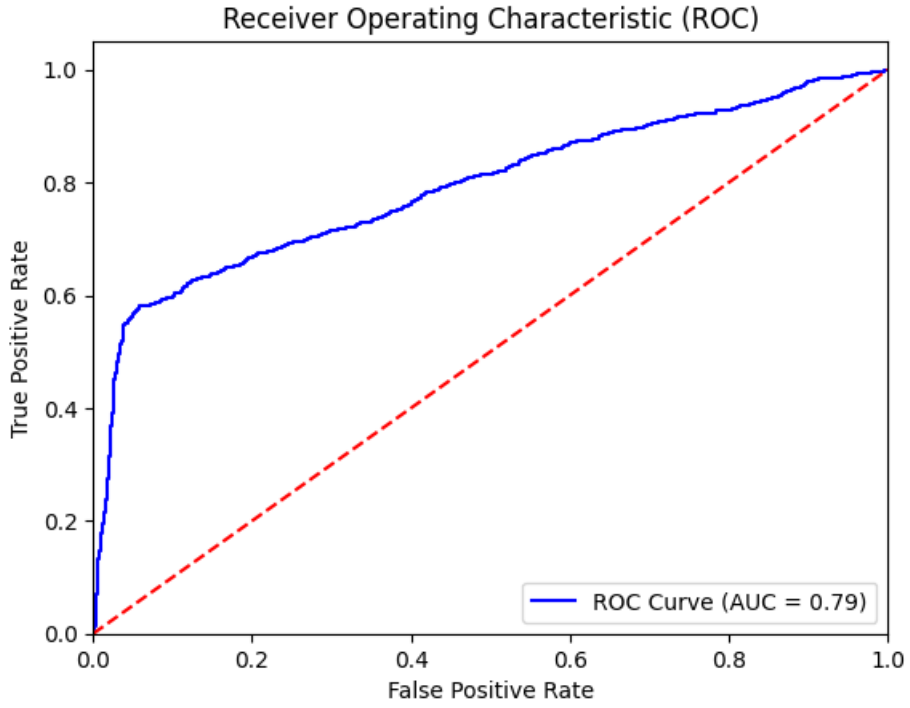
The area under the receiver operating characteristic curve (AUC) is an industry-standard accuracy metric used for binary classification models. **AUC measures the ability of the model to predict a higher score for positive examples, as compared to negative examples.** The AUC metric provides an aggregated measure of the model performance across all possible classification thresholds. It is not dependent on the choice of a specific threshold value used to map the probabilities predicted by a model into positive and negative classifications.

The AUC metric returns a decimal value from zero (0) to one (1). AUC values near 1 indicate an ML model that is highly accurate. Values near 0.5 indicate an ML model that is no better than guessing at random. Values near 0 are unusual to see, and these typically indicate a problem with the data.

Essentially, an AUC near 0 says that the ML model has learned the correct patterns, but is using them to make predictions that are as inaccurate as possible (0s are predicted as 1s and vice versa). For more information about AUC, go to the Receiver operating characteristic (https://en.wikipedia.org/wiki/Receiver_operating_characteristic) page on Wikipedia.

A binary model that classifies no better than random guessing, with equal rates of true and false positives, has a AUC score of 0.5 . The curve representing a random binary classifier is a diagonal dotted red line in a receiver operating characteristic graph. The curves of more accurate classification models lie above this random baseline, where the rate of true positives exceeds the rate of false positives.

AUC of candidate **WeightedEnsemble_L2_FULL** is **0.79**.



Correct predictions

- **True positive (TP)**: *WeightedEnsemble_L2_FULL* predicted the value as 1 , and the true value is 1 .
- **True negative (TN)**: *WeightedEnsemble_L2_FULL* predicted the value as 0 , and the true value is 0 .

Erroneous predictions

- **False positive (FP)**: *WeightedEnsemble_L2_FULL* predicted the value as 1 , but the true value is 0 .
- **False negative (FN)**: *WeightedEnsemble_L2_FULL* predicted the value as 0 , but the true value is 1 .

False positive rate

The false positive rate (FPR) measures the false alarm rate or the fraction of actual negatives that were falsely predicted as positives. The range is 0 to 1. A smaller value indicates better predictive accuracy.

FPR = FP/(FP+TN)

True positive rate

The true positive rate (TPR) measures the fraction actual positives that were predicted as positives. The range is 0 to 1 . A larger value (1 being the largest) indicates better predictive accuracy.

TPR = TP/(TP+FN)

Precision recall curve

The **precision recall curve** demonstrates the trade-off between precision and recall. The higher the precision and recall, the larger the area under the curve. For more information, see Precision and recall (https://en.wikipedia.org/wiki/Precision_and_recall) in Wikipedia.

Precision

Precision measures the fraction of actual positives that are predicted as positive out of all those predicted as positive. The range is zero (0) to one (1). A larger value indicates better accuracy in the values predicted.

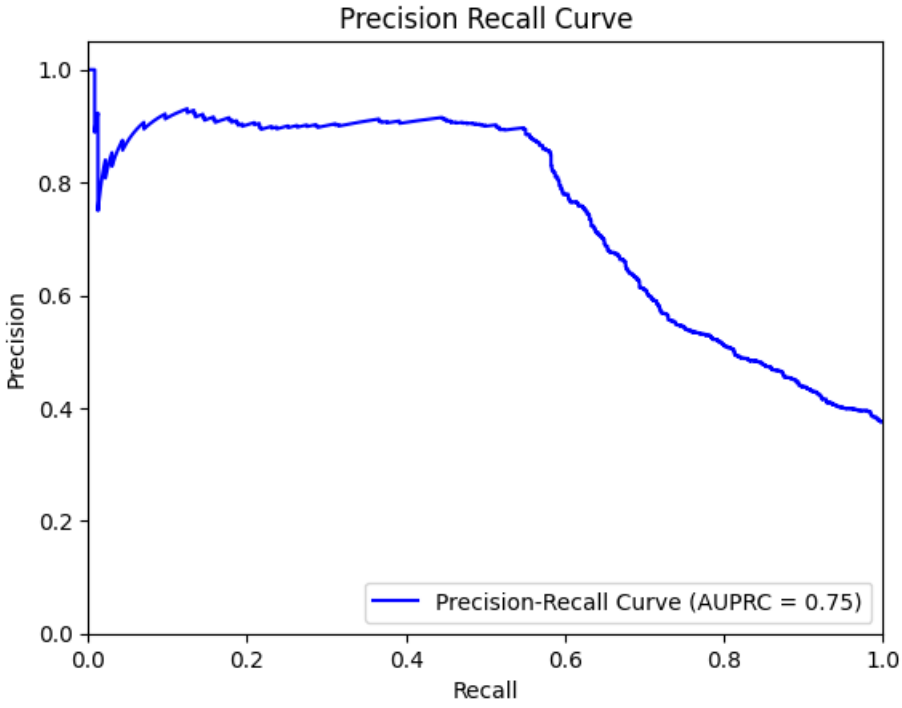
Precision = TP/(TP+FP)

Recall

Recall measures the fraction of actual positives that are predicted as positive out of all of the actual positives in the sample. This is also known as the sensitivity and as the true positive rate. The range is zero (0) to one (1). A larger value indicates better detection of positive values from the sample.

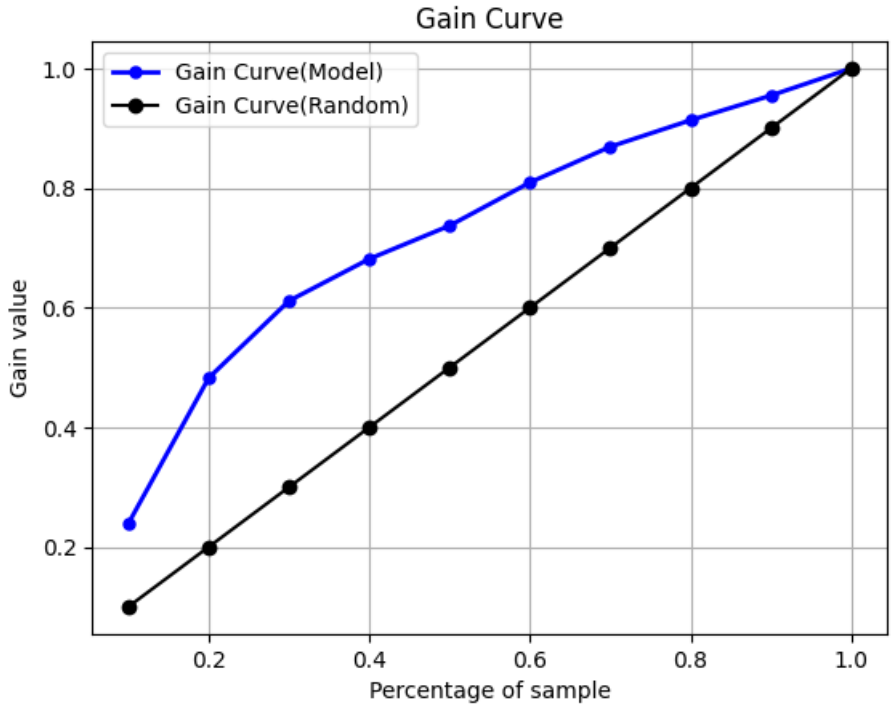
Recall = TP/(TP+FN)

The **WeightedEnsemble_L2_FULL** candidate has **precision** of **0.86** and **recall** of **0.58**.



Gain curve

The **gain curve** predicts the cumulative benefit of using a percentage of the dataset to find a positive label. The gain value is calculated during training using the ratio of the cumulative number of positive observations to the total number of positive observations in the data at each decile. If the classification model created during training is representative of the unseen data, you can use the gain curve to predict the percentage of data you would need to target in order to obtain a percentage of positive labels. The greater the percentage of the dataset used, the higher the percentage of positive labels found.



Lift curve

The **lift curve** illustrates the uplift of using a trained model to predict the likelihood of a finding positive label compared to a random guess. The lift value is calculated during training using the ratio of percentage gain to the ratio of positive

labels at each decile. If the model created during training is representative of the unseen data, you can use the lift curve to predict the benefit of using the model to target positive labels over randomly guessing.