Model Evaluation:

Model assessment is a method of analyzing a machine learning model's performance, as well as its advantages and limitations, using a variety of evaluation criteria. Model evaluation is essential for assessing the efficacy of a model during the first phases of research, as well as for model monitoring.

Boosting Overfit:

A boosting strategy is a machine learning method that predicts a target variable using a mixture of weak learners. Boosting is a method for increasing the number of weak learners in an ensemble by adding a new weak learner iteratively. As training data, the new weak learner employs the error of the ensemble from earlier rounds. This method continues until the ensemble reaches a predetermined performance level or the maximum number of iterations specified is reached.

**Supported Models**

We can examine the following models:

AdaBoost (sklearn) - GradientBoosting (sklearn) - XGBoost (xgboost) - LGBM (lightgbm) - CatBoost (catboost)

Goal: Examine for overfitting produced by a gradient boosted model with too many iterations.

Calibration Score

The Brier score can be used to evaluate the calibration of a classifier. For further details, please visit <https://en.wikipedia.org/wiki/Brier_score>

Useful link to read more:

<https://neptune.ai/blog/brier-score-and-model-calibration>

<https://machinelearningmastery.com/probability-calibration-for-imbalanced-classification/>

Calibration curves (sometimes referred to as reliability diagrams) compare the accuracy of the probabilistic predictions of a binary classifier. It compares the actual frequency of the positive label to its projected probability for binned forecasts.

Goal: Generate the calibration curve for each class using the brier score.

Confusion Matrix Report

What is a classification report?

As the name suggests, it is the report which explains everything about the classification. This is the summary of the quality of classification made by the constructed ML model. It comprises mainly 5 columns and (N+3) rows. The first column is the class label’s name and followed by Precision, Recall, F1-score, and Support. N rows are for N class labels and other three rows are for accuracy, macro average, and weighted average.

Precision: It is calculated with respect to the predicted values. For class-A, out of total predictions how many were really belong to class-A in actual dataset, is defined as the precision. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] column.

Recall: It is calculated with respect to the actual values in dataset. For class-A, out of total entries in dataset, how many were actually classified in class-A by the ML model, is defined as the recall. It is the ratio of [i][i] cell of confusion matrix and sum of the [i] row.

F1-score: It is the harmonic mean of precision and recall.

Support: It is the total entries of each class in the actual dataset. It is simply the sum of rows for every class-i.

Source: https://www.geeksforgeeks.org/compute-classification-report-and-confusion-matrix-in-python/

Goal : Compute the model's confusion matrix using the provided dataset.

Model Error Analysis:

Evaluating the model's overall performance metrics provides a valuable summary and can be used to track model progress during training or to compare models. However, when it's time to thoroughly assess whether a model is ready for production, or when you want a deeper understanding of your model's performance in order to enhance it or become aware of its flaws, it is advisable to examine how the model works on different parts of the data. The model error analysis check looks for data segments where the model error is much smaller than the model error of the entire dataset.

Algorithm:

1. Computes the loss per sample (for log-loss for classification, mse for regression).
2. Based on the input features, trains a regression model to forecast the error of the user's model.
3. Repeat step 2 multiple times with different tree parameters and random states to ensure that the most pertinent partitions for model error are chosen.
4. The features with the highest feature relevance score for the error regression model are chosen, and the error distribution is shown against the feature values.

Goal: Determine the characteristics that best segment the data into high and low model error segments.

Model Inference Time:

Goal: Determine the model's average inference time per sample in seconds.

Model Info:

Goal : Summarize given model parameters.

Multi Model Performance Report:

Goal: Summarize the performance evaluations of several models on testing datasets.

Performance Report:

Goal: Compile the results of a dataset and a model.

ROC Report:

A ROC curve (receiver operating characteristic curve) is a graph that illustrates the performance of a classification model over all categorization levels. This graph illustrates two parameters: Rate of True Positivity Rate of False Positives.

Useful Source:

https://en.wikipedia.org/wiki/Receiver\_operating\_characteristic

Goal : Determine the ROC curve for each classification.

Segment Performance:

Goal: Present performance score as a heatmap split by two top (or specified) attributes.

Train Testing Prediction Drift:

Goal: Using statistical metrics, compute the prediction drift between the training dataset and the testing dataset.

Unused Features:

How do underused features impact my design?

Due to "The Curse of Dimensionality" or "Hughes Phenomenon," having too many features can increase training times and reduce model performance. This is because the number of features exponentially increases the dimensional space. When the space is too big relative to the quantity of data samples, the samples are distributed in an extremely sparse manner. This sparsity also makes the samples more similar to one another because they are all far from one another, making it more difficult to cluster comparable samples in order to identify patterns. Increased sample similarity and dimensionality may need more sophisticated models, which are more susceptible to overfitting.

Features with low model contribution (feature significance) are likely noise and should be eliminated because they increase the dimension without contributing. However, models may lack essential features. In light of this, the Unused Features check eliminates features with high variance, as they may represent data that was disregarded during model creation. We may desire to carefully verify these characteristics to ensure that our model does not exclude vital information.

Goal: Identify largely unused features within the model.

Weak Segments Performance:

Goal: Find segments with low performance ratings.