**Model Training Documentations Updated**

1. **Confusion Matrix**

The confusion matrix is a tool for summarizing the performance of a classification algorithm. A confusion matrix will give us a clear picture of classification model performance and the types of errors produced by the model. It gives us a summary of correct and incorrect predictions broken down by each category. The summary is represented in a tabular form.

Four types of outcomes are possible while evaluating a classification model performance. These four outcomes are described as follows:-

**True Positives (TP)** – True Positives occur when we predict an observation belongs to a certain class and the observation actually belongs to that class.

**True Negatives (TN)** – True Negatives occur when we predict an observation does not belong to a certain class and the observation actually does not belong to that class.

**False Positives (FP)** – False Positives occur when we predict an observation belongs to a certain class but the observation actually does not belong to that class. This type of error is called **Type I error.**

**False Negatives (FN)** – False Negatives occur when we predict an observation does not belong to a certain class but the observation actually belongs to that class. This is a very serious error and it is called **Type II error.**

1. **ROC Curve**

A tool for measuring classification model performance visually in **ROC Curve**. ROC Curve is Receiver Operating Characteristics Curve. A ROC Curve is a plot which shows the performance of a classification model at various classification threshold levels.

The ROC Curve plots the **True Positive Rate (TPR)** against the **False Positive Rate (FPR)** at various threshold levels.

**True Positive Rate (TPR**) is also called **Recall**, it is defined as the ratio of **TP to (TP + FN)**

**False Positive Rate (FPR)** is defined, as the ratio of **FP to (FP + TN)**

**In ROC Curve**, we would focus on the TPR and FPR of a single point. This will give us the general performance of the ROC curve at various thresholds.

***A perfect class classifier of ROU AUC equals o 1, whereas a purely random classifier will have a ROC AUC equal to 0.5***

1. **F1 - Score**

F1-score is the weighted harmonic mean of precision and recall. The best possible f1-score would be 1.0 and the worst would be 0.0. So f1-score is always lower than accuracy measures as they embed precision and recall into their computation.

The weighted average of f1-score should be used to compare classifier models, not global accuracy.

1. **Classification Report**

**Classification report** is another way to evaluate the classification model performance. It displays the precision, recall, f1 and support scores for the model.

1. Model Evaluation and improvement

**There** a**re several techniques used to improve model performance, but i will be discussing some of the techniques here:**

1. **Recursive Feature Elimination with Cross Validation ¶**

Recursive feature elimination (RFE) is a feature selection technique that helps us to select the best features from the given number of features. At first, the model is built on all the given features. Then, it removes the least useful predictor and builds the model again. This process is repeated until all the unimportant features are removed from the model.

1. **Recursive Feature Elimination with Cross-Validated (RFECV) feature** selection technique selects the best subset of features for the estimator by removing 0 to N features iteratively using recursive feature elimination. Then it selects the best subset based on the accuracy or cross-validation score or roc-auc of the model. Recursive feature elimination technique eliminates n features from a model by fitting the model multiple times and at each step, removing the weakest features.
2. **Hyperparameter Tuning**

**Hyperparameter tuning makes** it possible for a model to have an increase in accuracy for effective predictions. Define a search space as a bounded domain of hyperparameter values and randomly sample points in that domain.

**Hyperparameters techniques are:**

**In a case of 10 million employees in a company with departments (A, B, C, D), and we were told to look for Micheal Edekin among all the 10 million employees, how do we create an effective model to do this:**

1. **Random Search CV ¶**

The best part about random search is that in the search for ***Michael Edekin*** Random Search narrows down our search to pick two of the four departments in which the needed employer should be found, giving that in Department A and D, the needed employer should be found.

**Department A = 72%**

**Department D = 53%**

While the other two departments would not be outlined for any relationship to the needed employer.

1. **Grid Search CV**

After which Random search has optimized our search to the possible minimum, then will Grid Search CV go to each department A and D and do a processed step by step deep search of each of the departments to find the employer.

In a case where ***Edekin Michael***  is in Department D while ***Michael Kings, Michael Owen, Michael Alabi*** are in Department A.

Grid Search then fetches out Department D and the name of the needed employer.

IT’S THAT EASY.