**UpTrain**

**Problem Statement**

● Finetune distilbert-base-uncased for the product and customer support team at Nike.

● Text = “Nike shoes are very [Mask].”

○ What you want: “Comfortable”, “Durable”, “Good-looking”

○ What you don’t want: “Expensive”, “Ugly”, “Heavy”, “Dirty”

● Find (or create) data-sources to fine-tune your model upon

● Define edge-cases to filter out just the necessary data

● Leverage statistical tools to visualize model health, do data cleaning etc.

**Model accuracy :**

To calculate the accuracy of the model on a given dataset, you can use the accuracy\_score function from the sklearn.metrics library. The function takes as input the true labels (y\_test) and the predicted labels (y\_pred) and returns the accuracy score as a decimal value between 0 and 1.

Here's an example:

accuracy = accuracy\_score(y\_test, y\_pred)

print('Accuracy:', accuracy)

The accuracy score is just one metric to evaluate the performance of a classifier. There are many other metrics such as precision, recall, and F1-score that may be more appropriate, depending on the specific problem and requirements.

**Edge cases defined for the task :**

An edge case is a**problem or situation that occurs only at an extreme (maximum or minimum) operating parameter**.

For example, a stereo speaker might noticeably distort audio when played at maximum volume, even in the absence of any other extreme setting or condition. An edge case can be expected or unexpected.

**Statistical measures:**

There are several statistical measures used in machine learning tasks to evaluate the performance of models and algorithms:

**Accuracy:** It measures the proportion of correct predictions made by the model.

**Precision**: It measures the number of true positive predictions out of all positive predictions made by the model.

**Recall (Sensitivity**): It measures the number of true positive predictions out of all actual positive instances.

**F1 Score:** It is the harmonic mean of precision and recall, and provides a balance between them.

**Confusion Matrix**: A matrix representation of the true positive, false positive, true negative, and false negative predictions made by a model.

**ROC (Receiver Operating Characteristic) Curve:** It is a graphical representation of the performance of a binary classification model.

**AUC (Area Under the ROC Curve):** It measures the performance of a binary classification model, where higher values indicate better performance.

**Log Loss:** It measures the error of a classification model that outputs probabilities for each class.

**Quality of Data:**

The quality of data sources used in a machine learning task plays a crucial role in the performance of the model. A few factors that determine the quality of data sources are:

**Relevance:** The data must be relevant to the problem being solved and should contain information that can be used to train the model.

**Completeness:** The data must be complete, meaning that it should contain all relevant information needed to train the model. Missing data can lead to inaccuracies in the model.

**Consistency:** The data must be consistent, meaning that the same information should be recorded in the same way across all data points. Inconsistent data can lead to inaccuracies in the model.

**Accuracy:** The data must be accurate, meaning that the information recorded should be correct. Inaccurate data can lead to inaccuracies in the model.

**Timeliness:** The data must be timely, meaning that it should be up-to-date and reflect the current state of the problem. Outdated data can lead to inaccuracies in the model.

**Formatting:** The data must be in a format that can be processed by the machine learning algorithms. Poorly formatted data can lead to difficulties in processing and modeling.

Ensuring the quality of the data sources used in a machine learning task is important in order to train accurate models and make reliable predictions.