# **LightGBM PROOF OF CONCEPT**

## **PREPARATION**

* Problem definition: Current text classification models [suffer in performance](https://nlp.stanford.edu/IR-book/html/htmledition/issues-in-the-classification-of-text-documents-1.html) due to a large number of similar text categories and are inefficient resulting in a higher carbon footprint. The LightGBM model can improve categorization by using internal features to easily communicate with the end user, resulting in reduced training time. LightGBM uses a histogram-based algorithm to reduce memory thus contributing to Green AI efforts. The deliverables are an end-to-end run, model, code and dataset.
* Data Selection/Preparation: There are four binary [datasets](https://ml.azure.com/dataset/lightgbm_binary/latest/details?wsid=/subscriptions/6560575d-fa06-4e7d-95fb-f962e74efd7a/resourceGroups/UW-Embeddings/providers/Microsoft.MachineLearningServices/workspaces/Embeddings&tid=72f988bf-86f1-41af-91ab-2d7cd011db47); 2 training and 2 testing datasets. [List of parameters and methods](https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.Dataset.html)

## **DEVELOPMENT**

* The text classification [model](https://ml.azure.com/model/sk-learn-random-forest-reg-model:1/details?wsid=/subscriptions/6560575d-fa06-4e7d-95fb-f962e74efd7a/resourceGroups/UW-Embeddings/providers/Microsoft.MachineLearningServices/workspaces/Embeddings&tid=72f988bf-86f1-41af-91ab-2d7cd011db47) uses an efficient histogram-based algorithm. Training with LightGBM has high improvement rates in speed. Most times when efficiency is increased accuracy can decrease, but not in this case. There is a small increase in [AUC metric](https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/#:~:text=There%20has%20been%20only%20a,when%20dealing%20with%20large%20datasets.) when compared with XGBOOST. The downside is the small user base regarding documentation and setup may be more involved than requirements for GPU. [LightGBM](https://lightgbm.readthedocs.io/en/latest/Features.html) grows trees leaf-wise instead of depth-wise which can cause over-fitting, but this is reduced by using a max-depth parameter.The user base is increasing so the tradeoff is well worth it.
* Collaboration: The features are customized for the end user allowing rapid iteration and validation of the trained models.
* Testing: The model is efficient and delivers in a short amount of time with notes for end users on how to make adjustments. More testing is needed to determine performance in a production setting. New tests should include new datasets not provided by the LightGBM example.

## **VALIDATION**

Initially, it was not expected for experiments to complete in less than 3 minutes [if the environment is configured correctly]. Will need to be tested with a larger never seen before dataset. Recommend using MLflow auto logging to keep track of metrics due to the ease of implementation.

## **METRICS**











