Experiment Design: Reducing Prediction Churn

The experiments' purpose is to find the best way for reducing churn. For our experiment, we will compare distillation and anchor techniques. We will also compare them to other metrics and datasets, but taking into account that we will be approaching this experiment from a classification standpoint.

Infrastructure: Data bricks, Azure, ML Flow, Auto ML

Alternative: Google Collab pro

The working environment for the experiment will be Colab using a GPU. To register the experiment, we will use MLFlow.

Metrics

We have chosen multiple metrics to compare the two methods over. A larger number of metrics allows us to observe more detailed differences in the methods while potentially highlighting pros/cons or trade offs between the different methods.

- **Churn:**

  $$C(f_1, f_2) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}_{f_1(X)f_2(X)<0}]$$

  Defined for the binary setting here. Here, f_1 is the newer model, f_2 is the older model, D is the dataset. Churn measures the expected amount of disagreements between two models.

- **Churn Ratio:**
  Let the old model be f_0, new model trained with a methodology be f_1, and new model trained without any methodology be f_2. Churn ratio is defined as $C(f_0,f_1) / C(f_0,f_2)$.

- **Win loss Ratio (WLR):**
  A **Win** is when the new model is able to correctly classify data points which the old model was incorrectly classifying. A **Loss** is when the new model incorrectly classifies data points that the old model was correctly classifying. WLR is the ratio between these and should be interpreted along with the p-value.

- **Accuracy:**
  Simple accuracy of the newly trained model.

- **Good Churn and Bad Churn:**
  Split churn into two categories: good churn - when a teacher model's predictions are incorrect and a student model predicts correct classes; bad churn - when a teacher model's predictions are correct but a student model predicts incorrect classes

Datasets

Cifar10 (images):

The CIFAR-10 dataset consists of 60000 32x32 color images in 10 classes, with 6000 images per class.

Online News popularity (tabular):

39797 observations, 61 columns

IMDB (text):

50000 total.

	Cifar10 (10 classes)	Online News popularity (2 classes)	Imdb (2 classes)
Total train + val size	50000	30000	40000
Total test size	10000	10000	5000
Old model train size	30000	15000	30000
New model train size	40000	20000	40000
Validation size	10000	9797	5000

Training Methodology

A unit of experiment will be a unique combination of label modification methodology, model size, dataset, and methodology hyperparameters.

Pre-processing technique: For each dataset the pre-processing pipeline will be standardized for the distillation and anchor methods.

Architecture: We will fix one architecture size for each dataset to experiment on. In the future, we want to experiment with different sizes. For example, on the Cifar10 dataset we can try out Resnet18 (11689512 parameters), Resnet50, and Resnet101.

Methods

- Knowledge Distillation
- Anchor
- Baseline (no label modification)

Model Training Hyperparameters: These hyperparameters are the traditional Deep Learning hyperparameters such as learning rate, batchsize, dropout rate etc. We will set up an automated approach to tune model training hyperparameters that can be objectively applied to each experiment. We will use the architecture's suggested hyperparameters as a baseline and experiment with a predefined set of altered model training hyperparameters.

Methodology Hyperparameters: These hyperparameters are the hyperparameters controlling the label modification (eg. lambda in knowledge distillation). Each choice of methodology hyperparameter will be treated as a unit of experiment.

Evaluation Metrics: We will evaluate each experiment using the evaluation metrics as mentioned in the metrics section of this document.

Results

The example table below summarizes our experimental approaches, including data set comparisons, methodology hyperparameters, and evaluation metrics.

		Images				**Text**				**Tabular**			
		Churn Ratio	**WLR**	**Good churn**	**Bad churn**	**Churn Ratio**	**WLR**	**Good churn**	**Bad churn**	**Churn Ratio**	**WLR**	**Good churn**	**Bad churn**
Distillation	lambda = 0.1												
	lambda = 0.2												
	lambda = 0.8												
Anchor	alpha=0.8, epsilon=0.1												
	alpha=0.8, epsilon=0.2												