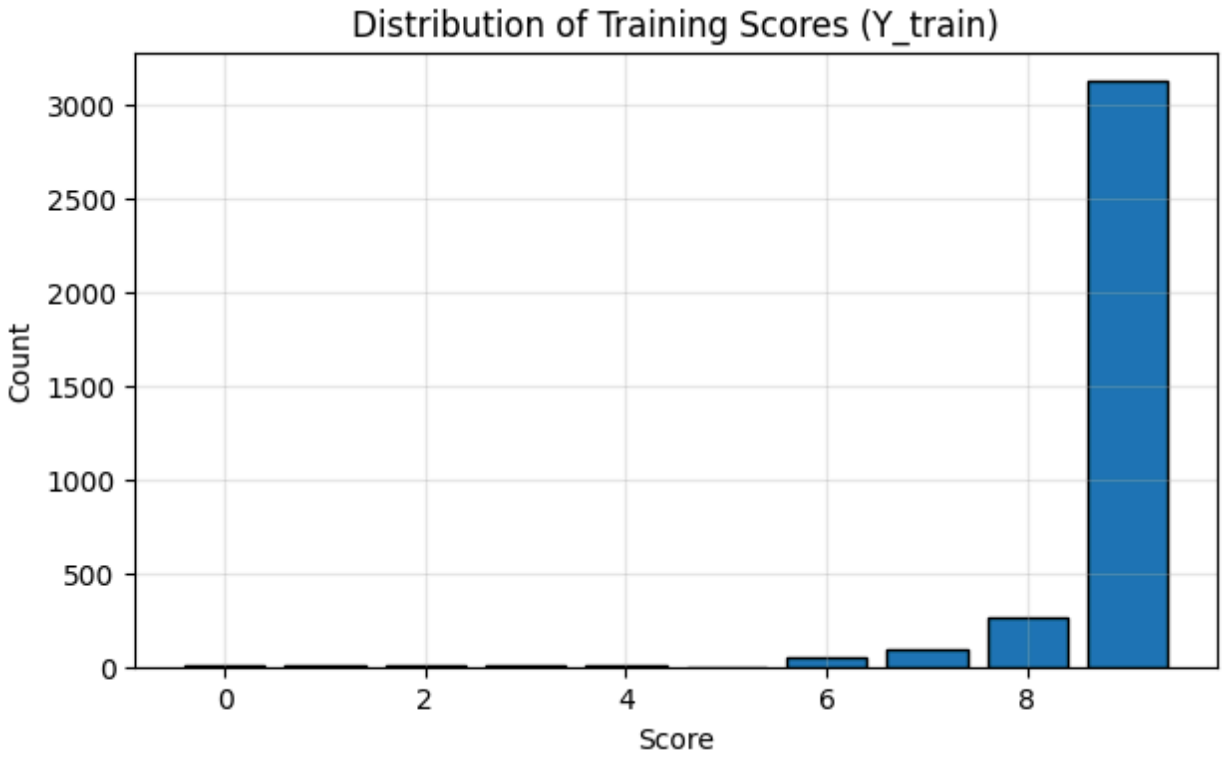
Kaggle contest Report

Problem Overview

The task is to predict a 0-10 (whole number) score between an AI evaluation metric and the corresponding prompt response pair. We are only provided the embeddings for the metric definitions.

The main challenge is that the training data is heavily skewed towards the 8-10 range.



Embeddings

I created embeddings using 2 methods. One is with paraphrase-multilingual-MiniLM-L12-v2 and the second is with the same model used by Prof in creating the metric definition embeddings. I proceeded to not use the 2nd set as it apparently does not translate all languages. (Perhaps it was fine for definitions because it may all have been in English.)

Method-1 Baseline XGBoost

I initially tried the XGBoost regressor. It gave an error of 4.6. When I looked at the predicted values, all of them were in the range of 8-10 only. Hence, to correct this bias, I introduced sample weights by dividing the training data into bins from the y-label values and using 1/freq as weights. This improved the RMSE significantly to 3.7.

Method-2 Siamese Regressor

I tried many different models. It didn't seem to improve much. Siamese regressor was able to reduce it to 3.6.

This showed that non-linear metric learning helped.

Method-3 Negative resampling- Final best csv

After some research I found a method called negative resampling wherein we take entries that have high score 9-10, swap the metric with another metric and give it a low score (0). This creates contrastive pairs so the model can learn what a mistake looks like.

But XGBoost did not work well with this model. This is because tree models struggle with semantic embeddings similarity.

I tried an MLP regressor with this negative resampling, and it reduced the RMSE to 2.7, which was my final and best submission.