**Machine Learning (2150534602) Term Project Report**

20182691 Wonseok Do

**Summary:**

**Introduction:**

중간 과제로 제출한 Term Project에는 몇가지 이슈가 있었다. 이번 Final Team 프로젝트에서는 마이너 이슈들을 처리하고 전보다 간단한 접근법을 사용해보려 한다. 기본적으로 문제의 퀄리티는 평가단이 결정한다. 어떤 평가원은 오른쪽이 좋다고 할 수도, 다른 평가원은 왼쪽이 좋다고 말할 수 있다. 평가 기준은 어느정도 공개되어 있으나, 결국 개인의 주관이 들어가기 때문에 쉽게 예측할 수 있는 문제는 아니다. 그렇기에 보유한 데이터에서 문제의 퀄리티에 영향을 줄 만한 요소들을 결정해 그것들을 기반으로 퀄리티를 예측하게 한다.

**Methods:**

I have assumed that all 'Wrong' answers are hard questions.

Confidence in Answer Metadata is the Percentage confidence score given for the answer. 0 means a random guess, 100 means total confidence. However, there were a lot of Missing Values in this data. Even though the values were virtually unusable, I thought the students' Confidence numbers were an indirect indication of the difficulty or clarity of the question. To fill in this value, I thought I'd calculate the probability of appearance based on the distribution of Confidence.

I used a multiple linear regression approach to predict quality. To predict quality, I decided to use KMeans to cluster similar questions together and then assign a score to each cluster based on the characteristics within the cluster. Anyway, since I don't know exactly, I calculated the score by taking the average of the [Confidence] and [IsCorrect] percentages for each cluster. For each data point, I calculated a score based on a combination of [Confidence] and [IsCorrect], which I weighted by the score for the cluster as a whole to get a final quality score.

**Discussion:**

When I measured the null data without looking directly at the content of the data, I found that there was one null data in the Test Dataset. However, since the Test Dataset is a private judgment, I decided to ignore it.

Confidence indicates the student's confidence in the question, so it can be an easy question or a clear question. Since no one else can know the source of the confidence, only the person solving the problem, I'll assume that a high-confidence problem is a good problem.

**Conclusion:**

The idea of splitting it into clusters and splitting the order within each cluster seemed plausible at the time. However, I had doubts about the correlation of the values beforehand, and since it was an unsupervised learning method with no right answer, I had to question the whole process. In fact, if I had used a simpler method, I might have gotten higher accuracy, but I wanted to try a new method and didn't have time to tweak the hyperparameters, so I think there's room for improvement for now.

**References:**

1. Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., … Zhang, C. (2020). Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2007.12061
2. Wang, Zichao, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, et al. “Results and Insights from Diagnostic Questions: The Neurips 2020 Education Challenge.” PMLR, August 7, 2021. https://proceedings.mlr.press/v133/wang21a.html.