GE Machine Learning Assignment

Thanks for taking the time to complete GlobalEnglish's machine learning assignment!

The dataset you received is a modified version of the dataset offered in a Kaggle competition organized by The Hewlett Foundation in 2012, intended to find a way to effectively automate the grading of student-written essays.

In the dataset provided for this assignment, you will find the following fields:

- essay_id: a unique identifier for each individual student essay
- essay_set: 1-8, an id for each set of essays. Note that each of the eight data sets has its own unique characteristics.
- essay: the ascii text of a student's response
- rater1_domain1: a grade given by an expert
- rater2_domain1: a grade given by a second expert
- domain1_score: resolved score between the raters

The encoding of the text data is latin-1.

After conducting an insightful analysis of the data, your task is the following:

1) create models to predict the resolved score between raters (domain1_score) of the essays and comment your findings
2) discuss the choice of the metric you considered to assess your models
3) find the main factors affecting the final grade

I recommend working in Python for this assignment, and send your solution as a Jupyter Notebook, but if you're more comfortable working with R, that's alright!

You are free to use any open source module/package, pretrained model or external data source if needed.

Good luck!

Giving a voice to global talent.