Project Proposal

Title: Predicting students engagement using machine learning techniques

Description: The goal of this project is to understand what factors are good predictors of the course-level engagement (attention, effort, positive emotional engagement scale) both in traditional (in-person) and remote learning settings using machine learning techniques. The forms of engagement scales are created based on survey responses received by the students during remote learning in Spring of 2020 and during traditional learning between 2016 and 2018.

Data description: This project includes 1329 data samples for each engagement scale (Y) which represents the student population of undergraduate students recruited across four engineering majors and 19 separate classes at the sophomore and junior levels in both traditional (in-person classroom) and remote learning settings. Self-reported ethnicity was Asian (43.3%), Black (2.9%), Hispanic (3.1%), White (40.7%), Pacific-Islander (less than 1%), Native American (less than 1%), and Other (2.6%). A number of students were mixed race (6.7%) among which White-Asian was most common (73%)

- ***Input (X):*** Student information for example: class, quarter, student status, major, college start year, age, gender, country status, race, family income, family education, social status, major situation. Full list of input is given here (link).
- ***Output (Y):*** I am going to predict three scales namely, attention scale, effort scale, and positive emotional engagement.
- ***Data source information:*** IRB (Internal Review Board) approval (provide application/study number here) was obtained to recruit and survey undergraduate students for this study. All participation was voluntary, and students were informed that their survey responses would remain confidential. (Data is provided by my supervisor Professor Denise Wilson)

Proposed approach outline: I will begin by wrangling data and prepare it for training by removing any duplicates, errors, missing values, normalize data, etc) and then visualize data to find the relationship between variables. After that I will split the data into training, validation and test sets. Next step will be to choose a regression model which can predict better results than a baseline. As a sanity check, I will be using MSE for evaluation of the model and will represent the engagement by plotting graphs. I will also be regularizing the model and will tune the parameters for improved performance.

Baseline solution: As a baseline, I will be using an OLS for comparison which will be less likely to overfit and will provide me a sense to go for a complex modelling.

Evaluation criteria: My evaluation criteria will be based on the root mean square error for the test data. I will also draw a graph plot of MSE for both validation and test data in order to detect overfitting.

Software: I will be using Jupyter Notebook to implement the model for this project using python language.