CSDA1050 Advanced Analytics Capstone Course

Project proposal

Improving student's graduation in Education

By Sylvain Kamto

Student ID: 11060

1- Introduction/ Background

Currently less than 65% of the students complete their studies as planned. Part of the students will move to work without graduation or change the branch of studies to another institute, but too many have either delayed in their studies (12.3%) or will completely discontinue (8.5%)

The delayed and dropout students pose significant direct costs to cities and schools due to reduced funding from government. Dropouts especially have challenges in finding a job and this problem is causing serious impacts on society in the long run.

To alleviate this problem, we are here by initiating a concept project on how to apply analytics to improve graduation in schools. The core of the idea is the following: utilize advanced analytics and machine learning to identify students who have elevated risk to dropout or delay in studies, so that interventions and support actions can be initiated early enough.

2- Research Question

Predicts which students have elevated risk of delayed studies or even dropping out

3- Dataset

Data were collected from the Office of Elementary and Secondary Education (OESE) of the U.S. Department of Education, state education agency (SEA) SIG applications, and SEA Websites.

The database contains 15,518 SIG-eligible schools across 50 states, the District of Columbia, and the Bureau of Indian Education (BIE), including 1,247 SIG-awarded schools across 49 states, the District of Columbia, and BIE.1

https://catalog.data.gov/dataset/ed-grants-school-improvement-grant-sig

or

https://www.ed.gov/sites/default/files/sig_database.xls

4- Dataset Description link

https://www.ed.gov/sites/default/files/SIG%20Database%20Documentation.pdf

5- Methodology

- Python Notebook will be used for codebase and analytics
- Dataset will need cleaning
- For the rating analysis and prediction, we explore several machine learning methods including Decision Tree, Random Forest, Support Vector Machine and Logistic Regression are considered to make relevant predictions.

6- Project deliverables timeline:

- Project Proposal – July 29, 2019
- Sprint #1 – Data Collection and exploration – July 29, 2019
- Sprint # 2 – codebase, report (brief), analysis plan - August 12, 2019
- Presentation review – August 20, 2019
- Final Project Submission – Final report, GitHub Repo, codes/analysis/results - August 27, 2019