Capstone Project Proposal

Gwen Rino

8/27/2017

# The Problem

Education is a staggeringly large project. In the United States alone, over 55 million children are enrolled in grades preK-12, and over 20 million adults attend American colleges and universities (source: National Center for Educational Statistics, <https://nces.ed.gov/>). Millions more nontraditional students are enrolled in university-sponsored MOOCs and other online programs, such as those offered by Coursera, edX, Khan Academy, and Springboard. The sheer volume of learning going on every day is heartening to all who recognize education as the single best pathway to the full realization of a person's human potential.

However, the project of education fails for too many students. Numerous reports show that over a million students drop out of American high schools each year, and nearly half of American college students do not complete a degree within six years. The cost to these individuals in lost opportunity, and the societal cost of increased poverty and alienation, are profound concerns.

# The Client

Educational institutions and programs need to be able to identify those students most at risk of failure early enough to intervene with appropriate support and remediation. Data analysis can predict which students would benefit from this extra consideration, greatly increasing learning, graduation rates, and, ultimately, human well-being.

# The Data

The University of California, Irvine, Machine Learning Repository includes a student performance dataset collected at two secondary schools in Portugal in 2008. It records students' performance in math and/or Portuguese, and also thirty other attributes that vary from mother's level of education to weekend alcohol consumption. The dataset can be found at <http://archive.ics.uci.edu/ml/datasets/Student+Performance>.

Citation: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

# Approach

After cleaning and tidying the data, I will do a series of exploratory visualizations and statistical analyses to determine which attributes are most strongly predictive of student performance. These will inform a machine learning program that can be applied to the dataset.

# Deliverables

The final project will include R code for the analysis and also a slide deck that presents the results.