Harvard E139 Project Proposal Fall 2015

Group Name: College_Score_Card_E139-Fall2015

Group Members:

Himanshu Dave	hdave@g.harvard.edu
Michael Prassanna Antony Raj	michael83@gmail.com
Swayambikash Panda	swayambikash@gmail.com
Bin Xiao	bin.xiao@wustl.edu

Github Link: *https://github.com/aspdave/College_Score_Card_E139-Fall2015*

Project Outline :

The expected outcome of this project is to provide deep insights into relationship between college ROI and its impact factors so that students can select colleges that are best fit for their needs.

The potential use cases could be: (1) the top schools students can refer to when choosing colleges, that can be achieved through creating a statistical model containing completion rate, earnings and tuition cost to determine best schools; (2) the relationship between college ROI and potential predictor variables such as SAT score, admission rate and so on (we can also analysis the relationship over time series), that can be achieved through creating regression model.

Our investigation is limited by Federal dataset available online. Time series is from 1996 -2013 USA colleges. (*https://collegescorecard.ed.gov/data/*).

Accomplishment as of today

Project group is formed and the kick-off meeting is done. Group's private share folder is created, group blog is created to share ideas and GitHub code share is established. Initial college scorecard dataset is data set is downloaded. We also have roughly examined the dataset and investigated the possible predicator variable(s).

What left to do

First, we will need to further examine the dataset and investigated data set for possible predictor variable and response variable. Second, we will examine the assumptions of the model: normality, independence and potentially constant variable. Third, we will build regression models between predictor variable(s) and response variable. Fourth, we will also generate some charts/graphics to better illustrate our findings. Finally, we will write up our project paper.

Challenges ahead

Defining the "Quality " and "true" cost factor of college is a real challenge. There is wide variety of variables to choose from. Although we do expect close to normal distribution due to large collection, there are good chances of outliers. We also need to check assumptions and limitations solely based on data contain with dataset. Cohort definition as imperfect and varies by each metrics.

Model does not fit all of the population as students tend to have different Goals and individual expectation that varied widely.