**Model to estimate the earnings potential of a student**

**Introduction:**

Taking a decision to attend a college is one of the most difficult decisions an individual has to make in their lifetime. Several factors that dictate this decision include cost of education, programs offered, the student debt incurred, earning potential after the completion of degree and others. Based on the personal preferences and financial condition, future students can have several colleges to choose from. There are websites that provide information on rankings of the colleges based on these criterions.

In this project, I want to focus on developing a model to estimate the earning potential of college graduates after completion of their degrees. I will also compare the salaries estimated by my model using the college scorecard data against the salaries provided by other available sources (e.g. websites).

Several variables are available from 1996 to 2016 for institutions of the United States including cost, graduation rate, debt, programs, SAT scores, repayment, earnings after school and others variables (<https://collegescorecard.ed.gov/data/>). A report from federal government compared the effect of these variables on the earnings of the students using aggregate and individual level regression approaches (Executive office of the President of the United States, 2015). In this project they also explored the effect of these variables on the earning potentials of the students.

Overall the objective of this capstone project is to:

1. Provide an exploratory analysis on the effect of independent variables on the earnings of the students.
2. To predict the earnings of a student graduating from a particular college using college score data provided by Federal government.
3. Compare the model results with the earnings of the college graduates from other websites such as www.payscale.com that provide salary reports for the best colleges by state.

**Dataset availability:**

1. College score dataset in the form of csv files downloaded from: <https://collegescorecard.ed.gov/data/>
2. Earnings of various alumni from different colleges from websites:

e.g. <https://www.payscale.com/>

**Potential beneficiaries:**

The model developed during this capstone project can help students to make better decisions that can maximize their earning potential, colleges to improve on their spending in the right direction to help their future students to be more successful, policymakers to get better information to provide colleges with more support.

**Methodology:**

All the codes for this project will be developed using python language. I will start the project with data ingestion from the csv files, cleaning the data using data wrangling techniques. Next step, would be to perform exploratory analysis by analyzing the effect of independent variables on the dependent variable (earning). Calculate initial statistical inferences from the data samples.

Divide the data into training and testing datasets. Use the training dataset to develop a model for estimating the earnings of an individual based on the independent variables. Test the dataset using the rest of the testing dataset. Score the dataset based on the model performance during training and testing phase.

Compare the results from the model against other dataset obtained from the website.

**Deliverables:**

The deliverables for this project will include codes used for the exploratory analysis and model development, results in the form of jupyter notebook and a report presenting the exploratory analysis, model, results, discussion and conclusions.

**References:**

Executive office of the President of the United States. (2015). Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education. Retrieved from https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf