Machine Learning Engineer Nanodegree

Capstone Proposal

Nathan Lile
January 6th, 2017

Proposal

Domain Background

In Western soscity, it is generally expected that students graduating from Highschool will go straight on to a 4 year university. However, with the cost of attending traditional 4 year universities rising steadily and student debt becomming a massive burden, second only to 30 year morgages, to those who choose to attend university, is it potentually wiser to wait several years before going on to University. This project will examin potentual corrilations between not going to university right out of highschool and obaining a more financually viable degree path. For me personally, this has personal baring since I fall into the catagory of a non-traditional student. I began my 4 year university track at age 25 after hafing worked construction and decided I was sick of that kind of work. I felt that having lived out in the world and really having a more solid understanding of the value and methods of things, I had a much more practical approach to university. In this project, I wanted to see if this was true for others.

Problem Statement

The problem I am working to solve is to draw a corrilation between Universities with higer rates of non-traditional students (students ages 25 - 65) and higher rates of graduation and higher rates of employment post graduation. This is in an effort to show that non-traditional students do a better job of choosing more viable degree paths and have a greater chance of finishing their 4 year degree. For the purposes of this model, a corrilation can be drawn if there is a better then 50% chance of schools with higher rates of non-traditional students also having porportionaly higher rates of graduation and employment post-graduation.

Datasets and Inputs

The goal is to draw corrilation from the College Scorecard dataset between schools with higher levels of non-traditional students (students who begin University from ages 25 to 65) and higher rates of graduation. Steps to be completed: - Download the College Score Card dataset(Data.gov) - Parse data into a usuable dataset with 4 or less metrics - Train clasifier on curated dataset - Attempt to infer a corrilation between higher levels of non-traditional students and better graduation rates and higher paying jobs out of school.

Solution Statement

A solution would be in the form of a classification “yes” or “no”. In this dataset, a “yes” would mean a greater then 50% (better then chance) corrilation between schools with higher numbers of non-traditional students, higher rates of graduation and higher paying positions straight out of university. Conversly, a “no” would be anything less then a 50% corilation.

Benchmark Model

At this time, I have been unable to find any benchmark models for this type of problem.

Evaluation Metrics

(approx. 1-2 paragraphs)

As an elevaluation metric, I will be comparing the percentage of the student population that are non-traditional with the rates of graduation. If the schools with higer numbers of non-traditional students also have roughly simmilar higher rates of graduation over schools with lower rates of non-traditional students then this metric should provide a clear measurment of better graduation rates among schools with higher concentrations of non-traditiional students.

Project Design

The project will follow a number of steps. - Parse our the relivent data in the data sets. Percentage of student body of schools that are age 25 and older. My work will focus on schools that have a high percentage of these students. Next I will parse the graduation rates at these schoolsand attempt to focus in on the graduation rates of non-traditional students if possible. last piece of information to parse will be the median incomes for graduates post university. - Once the information is parsed, I will seperate a training and testing data set. - Next, I will build a classifier, further research is needed to narrow the list of candidates for this step. - Once the classifier is working, I will utalize PCA and graphical models to view results the model is comming up with. This is to avoid the corrilation/causation slippery slope. - Schools that are classified as above 50% corilation will be written into a seperate list

---