

Sources - <https://github.com/caesar0301/awesome-public-datasets>

**Do non traditional students make more money with their degree then traditional students?**

**Is there a corrilation between not going to school right out of HS and more financually viable degree paths?**

## Machine Learning Engineer Nanodegree

### Capstone Proposal

Nathan Lile  
December 31st, 2050

### Proposal

*(approx. 2-3 pages)*

#### Domain Background

*(approx. 1-2 paragraphs)*

In this section, provide brief details on the background information of the domain from which the project is proposed. Historical information relevant to the project should be included. It should be clear how or why a problem in the domain can or should be solved. Related academic research should be appropriately cited in this section, including why that research is relevant. Additionally, a discussion of your personal motivation for investigating a particular problem in the domain is encouraged but not required. " In Western soccity, 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 mortgages, to those who choose to attend university, is it potentially wiser to wait several years before going on to University. This project will examin potential 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

*(approx. 1 paragraph)*

In this section, clearly describe the problem that is to be solved. The problem described should be well defined and should have at least one relevant potential solution. Additionally, describe the problem thoroughly such that it is clear that the problem is quantifiable (the problem can be expressed in mathematical or logical terms) , measurable (the problem can be measured by some metric and clearly observed), and replicable (the problem can be reproduced and occurs more than once).

### Datasets and Inputs *(approx. 2-3 paragraphs)*

In this section, the dataset(s) and/or input(s) being considered for the project should be thoroughly described, such as how they relate to the problem and why they should be used. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included with relevant references and citations as necessary It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem. " 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: 1) Download the College Score Card dataset(Data.gov) 2)

Parse data into a usable dataset with 4 or less metrics 3) Train classifier on curated dataset 4) Attempt to infer a correlation between higher levels of non-traditional students and better graduation rates and higher paying jobs out of school.

## Solution Statement

*(approx. 1 paragraph)*

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms), measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once). A solution would be in the form of a classification “yes” or “no”. In this dataset, a “yes” would mean a greater than 50% (better than chance) correlation between schools with higher numbers of non-traditional students, higher rates of graduation and higher paying positions straight out of university. Conversely, a “no” would be anything less than a 50% correlation.

## Benchmark Model

*(approximately 1-2 paragraphs)*

In this section, provide the details for a benchmark model or result that relates to the domain, problem statement, and intended solution. Ideally, the benchmark model or result contextualizes existing methods or known information in the domain and problem given, which could then be objectively compared to the solution. Describe how the benchmark model or result is measurable (can be measured by some metric and clearly observed) with thorough detail. At this time, I have been unable to find any benchmark models for this type of problem. ### Evaluation Metrics *(approx. 1-2 paragraphs)*

In this section, propose at least one evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model. The evaluation metric(s) you propose should be appropriate given the context of the data, the problem statement, and the intended solution. Describe how the evaluation metric(s) are derived and provide an example of their mathematical representations (if applicable). Complex evaluation metrics should be clearly defined and quantifiable (can be expressed in mathematical or logical terms).

## Project Design

*(approx. 1 page)*

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project. The project will follow a number of steps. 1) Parse out the relevant 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 schools and 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. 2) Once the information is parsed, I will separate a training and testing data set. 3) Next, I will build a classifier, further research is needed to narrow the list of candidates for this step. 4) Once the classifier is working, I will utilize PCA and graphical models to view results the model is coming up with. This is to avoid the correlation/causation slippery slope. 5) Schools that are classified as above 50% correlation will be written into a separate list

---

## Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?

- Are all the resources used for this project correctly cited and referenced?