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 correlations between various metrics that universities measure (age, gender, parent income etc) and see how strongly these effect graduation rates. As a machine learning paoblem, I will be training a model to predict the graduation rates of a given university given various data metrics collected from the student body. As a secondary focus, I will be examining if returning to University after several years in the real world has a strong positive effect on graduation rates. 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. In the real world, this information is usuful to educational institutations, parents, students and financual aid institutations alike. Knowing what factors are most important in effecting higher graduation rates among university students can help provide focused attention and support in these area to provide maximum support to the student body. For example, if the income of the parents and if the parents of the student attended university are large determinate, this could allow for more funding and support for low income and first generation students attendiing university.

NEED TO DO: -Please be sure to provide some background on your project. How have researchers approached this type of problem in the past. Can you cite a few examples in the introduction section? https://arxiv.org/pdf/1606.06364.pdf https://arxiv.org/pdf/1405.3727.pdf

Problem Statement

NEED TO DO: -This is the major problem with this proposal. You've picked a problem where the solution itself won't be measurable or replicable. For instance, let's say you find a correlation between non-traditional students and better post-graduation outcomes. Then what? How do you repeat this result? Is there a way to benchmark this task?

Here's an example of a possible way to structure the problem that might be more feasible. You could look at a number of factors about the student body such as age, gender, educational background, work experience, parental status etc. You then take a regression approach that tries to predict the graduation rate for each school based on these features.

After you've created a strong regression algorithm that performs well (this would be the main 'problem'), you could also examine the feature importances and see if a school's non-traditional student population is an important feature in predicting graduation rates.

Datasets and Inputs

The College Scorecard project is designed to increase transparency, putting the power in the hands of students and families to compare colleges and see how well schools are preparing their students to be successful. This project provides more data than ever before to help students and families compare college costs and outcomes as they weigh the tradeoffs of different colleges, accounting for their own needs and educational goals. These data are provided through federal reporting from institutions, data on federal financial aid, and tax information. These data provide insights into the performance of schools that receive federal financial aid dollars, and the outcomes of the students of those schools (CollegeScoreCard). https://catalog.data.gov/dataset/college-scorecard

NEED TO DO: -Specifically mention and describe the features that you're going to use from the dataset. If you take the regression approach that I suggest above, you may end up wanting to use a lot of these. You should describe them for the reader so that it's clear that you know what you want to do.

Solution Statement

NEED TO DO:

-In this section of the proposal, you should be describing the machine learning techniques that you'll use to solve the problem. It should be clear if you're using classification/regression techniques and you should specifically mention the machine learning algorithms that you plan on using (or at least trying initially). It's OK to have a list of things that you want to initially try (because you can't know a priori which methods will work the best).

Benchmark Model

NEED TO DO:

For my benchmark Model, I will be using a linear Support Vector Regressor from the SKlearn library. I am choosing this model for its grneral flexability and ability to scale to larger numbers of featurers. For the most part, I will be running this model out of the box, only using Gridsearch CV to optimise the meta-peramaters within the linear SVR

Evaluation Metrics

NEED TO DO: Provide evaluation metric

Project Design

NEED TO DO: -Please be sure to structure this as a list of steps that describe the specific machine learning/preprocessing techniques that you'll use in tackling your problem. Don't skimp on the details here. This section is a place where you can bounce your initial ideas off of the reviewers. We can possibly save you a lot of time by helping you quickly narrow your approach down to a few things that are likely to work. Be sure to list any preprocessing and machine learning techniques that you want to try.
