Machine Learning Engineer Nanodegree

Capstone Proposal

Nathan Lile
January 15th, 2017

Proposal

Domain Background

In Western society, it is generally expected that students graduating from High School will go straight onto a 4 year university. However, with the cost of attending traditional 4 year universities rising steadily and student debt becoming 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 examine potential correlations between various metrics that universities measure (age, gender, parent income etc) and see how strongly these affect graduation rates. As a machine learning problem, 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 category of a non-traditional student. I began my 4 year university track at age 25 after having 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 useful to educational institutions, parents, students and financial aid institutions alike. Knowing what factors are most important in affecting 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 attending university.

This question has been tackled by Machine Learning experts before. Typically, the income of the student's family, race and gender have been focused on as determining factors of graduation. In the two papers below, scientists found a good baseline model then built at least three regression models (typically including Random Forests) and worked with a dataset that had been reduced to relatively small numbers of features (typically less than 20). https://arxiv.org/pdf/1606.06364.pdf https://arxiv.org/pdf/1405.3727.pdf One thing that I will be focusing on that does not seem to be getting much attention

is the age of the students when they begin university as a determining factor on graduation rates.

Problem Statement

The goal of this project will be to build a strong regression model to predict the graduation rates of various universities based on featured inputs (such as the cost to attend the school, demographic data of the students, etc) with the highest level of accuracy (comparing the model output to the actual graduation rates from the universities). Additionally, once the model has been optimized, I will attempt to select the top 5 "most important" features from my dataset in identifying if a student will graduate or not.

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). The entire CollegeScoreCard dataset it rather large consisting of information from Title IV recipients, or students who receive federal grants (CollegeScoreCard) from four and two year Universities in the United States who participate in Title IV programs from 1996 to 2013. https://catalog.data.gov/dataset/college-scorecard

In this data set I will be using the following features financial, ademic information of the universities themselves and demographic data of the student body to attempt to predict the graduation rates, on a scale of 0.0 (zero percent graduation) to 1 (100% graduation rate) of the universities: -Degree Type -Public/Private Nonprofit/Private For-Profit -Revenue/Cost of the School -Programs Offered by Type -Admission Rate -SAT and ACT Scores -Average Cost of Attendance,Tuition and Fees -Number of Undergraduate Students -Undergraduate Student Body by Race: -Undergraduate Students by Part-Time/Full-Time Status -Undergraduate Students by Family Income -Undergraduate Student Body by Age (amount of students age 25-65) -Share of First-Generation Students -Percentage of Pell Students -Cumulative Median Debt of students loans

Solution Statement

To begin with, I will begin by using three models: RandomForestRegressor optimized with GridsearchCV, Ridge Regression with Gridsearch CV and a Bayesian Regression with Gridsearch CV. I will compare the models R2 score, runtime and prediction time to determine the efficiency/accuracy tradeoff of the models. Depending how the models perform, I will attempt to pipeline them into parallel regressors and see if those perform better than the singular models. For pre-processing of the data, I will be looking back to the Boston-housing project that I did in the beginning of this Udacity course for guidance with data exploration, performance metric development and developing data metrics that will provide a good fit.

Benchmark Model

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

Evaluation Metrics

As a custom evaluation metric, I will be comparing the output from my model to the 150 Percent IPEDS Completion Rate (completion rates for first-time, full-time students who begin school in the fall semester and complete within 150 percent of the expected time to completion) as reported by the universities themselves (CollegeScoreCard) This metric is provided in the dataset and will be the metric that my model will be attempting to predict. To measure how my model is performing internally, I will be using mean absolute error (MAE) or Root-mean-square-error (RMSE). These are both statistical methods of calculating deviation of my model compared to the desired outcome metric

Project Design

The data set I am working with is already formatted as a CSV file, so I will not need to perform additional formatting. I will begin by isolating the features I mentioned above. Next, I plan on following similar techniques used in the Boston Housing project at the beginning of this course: I will first explore the data, making sure that my data makes sense. Next, I will work on pre-processing the data since it will be in a mix of ints, floats and strings. For simple yes/no strings, I will convert them to binary 1/0 . For string answers that have a spectrum, in similar fashion to the student intervention project, separate each separate

item into its own column and give it a value of 1. I will likely be relying on gridsearchCV and pipelining to optimize my models but will also use visual performance graphs and tables if one of the models is acting oddly.
