To College or Not to College: Prediction of Post Collegiate Earnings and Debts. An exploration of the College Scorecard Dataset

Nithyashri Govindarajan, Rakshitha K Bhat

Problem

Deciding to pursue higher education and choosing a college can be a daunting task. There are many factors that influence this process, and the consequences are life changing. As data driven decision making makes great strides in human lives, college decisions can also greatly benefit from analysis of credible data that reflect current trends in education as well as future job prospects. According to the U.S Department of Education, student borrowers owe more than $1 trillion and about seven million borrowers are in default. With the increasing cost of higher education as well as the accompanying rise of student debt, prospective students can greatly benefit from knowing about post collegiate earnings and the correlation of these earnings to the size, quality of education, location of college or any other factors, if any. Additionally, these could be a useful indicator for the recovery of loans. Our project aims at using machine learning tasks to do exactly this. The goals of this project are:

- Prediction of post-graduation earnings
- Prediction of debt
- Determining the most influential factors in this prediction

Our models can potentially fill in the gaps with current data and unearth interesting facts about the factors that actually determine earnings and debt. Matching tuition costs and average potential earnings and debt can help determine typical lengths of student loans.

Dataset

The dataset we intend to use for the purpose of the project is the College Scorecard Dataset. This public dataset can be downloaded at https://collegescorecard.ed.gov/data/.
College Scorecard provides publicly available dataset consisting of 1976 metrics for approximately 7700 degree-granting institutions from 1996 through 2016 [1]. These metrics include demographic data, test scores, family income data, data about the percentages of students in each major, financial aid information, debt and debt repayment values, earnings of alumni several years after graduation, percentage of students from each ethnicity, college completion rates and many more. The U.S Department of Education launched this data collection drive through tax submissions of federally aided students as a means to help students make more informed college decisions. The data has raised eyebrows because of the following caveats - it reports findings for only students with federal aid and has omitted about 700 colleges, particularly community colleges. Despite this, the data is useful to provide general albeit approximate trends to aid decision making for all stakeholders.

Experiments and Methodology:

The task is a regression task with two target variables - earnings and debt. But we setup our problem as two separate tasks - one for prediction of earnings (keeping debts as a feature) and one for prediction of debt (keeping earnings as a feature). Our analysis has the following steps and issues to address with the chosen dataset.

1) Data Preprocessing - Handling non-standard values and privacy suppressed data points which we intend to solve with mean imputation of values, regression imputation (after learning a regression model without these values) and multiple imputations (using a set of observations from a simple model) as mentioned in [4].
2) Measuring Feature Importance and Selection of Features - Analysing the contribution of each feature through various wrapper and heuristic models. At this point, we want to look at recursive feature elimination, forward selection of sequential features and statistical dependence filtering. XGboost[5] also provides the most important features, so this can also be looked at in combination with boosted tree regressors.
3) Regression Task - Using a linear regression to set a baseline, we want to experiment with all regression models: K nearest neighbours, decision trees, support vectors and neural networks. Initial dabbling with the data indicates non-linearities between features, thus influencing the choice of support vector regression techniques that allow kernel function flexibilities.

As an extension, the problem is setup as a multivariate regression problem predicting both earnings and debt at the same time. We think that this will provide more accurate insights since a combination of just predictions of earnings and debt is biased as it uses the other value in making a prediction.

All evaluations of the performances of the regressors will be done using Root Mean Squared Error (RMSE) because we have standard gold labels to compare against. We will use Cross Validation for hyperparameter optimizations. All our work is geared for a Python 2 implementation using SciKit Learn support functions.

Related Work and Novelty:

Using machine learning for predictive tasks for college earnings and debt has been explored due to the lack of credible encompassing data. From a social science perspective, work has been produced analyzing factors influencing college decisions - such as quality of education and earnings [6][7][8]. Since College Scorecard is a relatively new dataset launched in September 2015, not much analysis has been done on it. There have been a few exploratory analysis conducted on Kaggle[2] regarding the nature of the data, the number of colleges for which data is available across the years, median SAT scores across universities, medial earnings etc. Agrawal et. al have recently performed analysis on predicting the collegiate earnings and debts[3] using various regression methods. Our work builds on these foundations, where, in addition to predictions of earnings and debts, we analyze the factors that contribute most to the correct predictions to provide more insights into the data. We want to empirically arrive at the important features unlike using a predetermined set of features, purely non-categorical, as in paper [3] which makes assumptions about the relative lesser influence of these features. We also intend to chart the trends in earnings of alumni across various years as opposed to a single year of analysis (2011) as in [3]. This we believe will help unearth more concrete correlations (if any) across data and average the findings across the years. With multivariate regression predicting both earnings and debt, two factors that are highly correlated, we extend the scope of independent predictions of earnings and debt which may not provide complete information by themselves. Thus, our task is more challenging with a larger dimensional data and more data points thus increasing modelling and computational complexity in every step of the methodology laid out.

Collaboration Plan:

Nithyashri	Rakshitha
Perform data preprocessing using mean imputation of values and class of multiple imputations using values from a simple model.	Perform data preprocessing using regression imputation of values.
Analyze the importance of various features and select the best features using the recursive feature elimination and forward selection algorithm.	Analyze the importance of various features and select the best features using statistical dependence filtering and XGboost.
Performing the regression task using K Nearest Neighbours and Support Vector Machines.	Performing the regression task using Decision trees, Linear Regression and Neural Networks. Examine bagging and boosting regressors.
Visualization and analysis.	Visualization and analysis.

References:

[1] U.S. Department of Education. College Scorecard, 2015.

[2] https://www.kaggle.com/benhamner/d/kaggle/college-scorecard/exploring-the-us-college-scorecard-data.
https://www.kaggle.com/kaggle/college-scorecard.

[3] http://cs229.stanford.edu/proj2015/212_report.pdf

[4] Benjamin M. Marlin. Missing Data Problems in Machine Learning, 2008: University of Toronto.

[5] Jerome H Friedmann. Greedy Function Approximation: A Gradient Boosting Machine, 2001: Institute of Mathematical Statistics.

[6] Dominic J. Brewer, Eric R. Eide, Ronald G. Ehrenberg. Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings, 1999: The Journal of Human Resources.

[7] Estelle James, Nabeel Alsalam, Joseph C. Conaty, Duc-Le To. College Quality and Future Earnings: Where Should You Send Your Child to College?, 1989: The American Economic Review.

[8] Lewis C. Solmon. The Definition of College Quality and Its Impact on Earnings, 1975: The National Bureau of Economic Research.