College Results

An Exploration of Institutional Data

Avriana Allen | 6.12.2020

Introduction

One of the biggest challenges when it comes to picking a college is the question of loans. We often intuitively look to data to answer that question, generalizing the results of the graduates to ourselves. But is it possible to predict this more formally? And if so, what can we learn from the predictors?

The Data

To answer such a question requires, first and foremost, data. The College Scorecard, a semi-recent site provided by the government to help students find institutes of higher education, is built off of the most recent institutional information. Containing private and public four year and two-year schools, the open data set is nothing if not comprehensive.

I chose to work with the 5-year loan repayment rate, a choice which proved useful and problematic. Loan repayment is formally defined as "the fraction of borrowers at an institution who are not in default on their federal loans and who are making progress in paying them down." It was not clear if a higher or lower percentage was ideal for a school. This variable was also only available for federal loans

The Models

I used MSE along with accuracy (calculated from MSE divide by variance) across the various models.

Linear Models performed very poorly. I used three different sets of variables, with the kitchen sink model: the top 10 variables from the forest models, four variable I thought would do well, and the most strongly correlated variables. Each model resulted in an MSE higher than the variance, suggesting that the models were entirely unable to explain the data.

Support Vector Machines did better. I ran the models with a linear, polynomial and radial kernel. I was only using variables with a correlation to the response variable. The best set of models was the SVM with the linear kernel, but it's accuracy as 34.5% at best.

The **Ridge** and **Lasso** models improved on these results, their best score coming in a 28.3% during training. I once again only used the correlated variables, given issues I was having with missing data on the other variables. Still, as it performed better than the other models before it, I decided to carry it over to the selection stage.

I also used **Random Forests**, and **Boosted** models. Both of these had lower MSEs, with accuracy rates around 18% and 20%. I used only numeric variables to train the models, and given their strong performance carried them over to the selection stage.

Selection and Performance

Running the models on the selection data produced fairly clear results. Random forests and boosted models were at the top with every other mode register at a 22.9% error rate or higher.

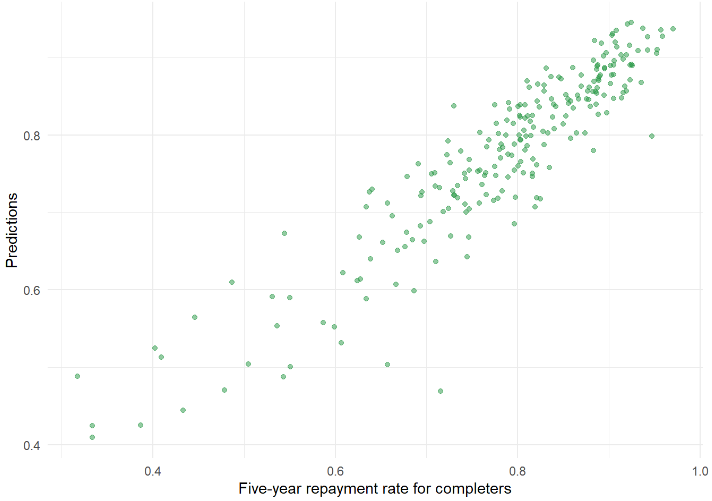
model	***mse***	***var***	***percent***
random_forest	0.00297	0.0181	0.165
boosted_3	0.00305	0.0181	0.169
ridge_min	0.00413	0.0181	0.229
ridge_lse	0.00420	0.0181	0.232
lasso_min	0.00424	0.0181	0.235
lasso_lse	0.00448	0.0181	0.248
boosted_2	0.01000	0.0181	0.555
boosted_1	0.04200	0.0181	2.330

Since the random forest model and the boosted model had very close results, I decide to carry them both over to the performance, where Random Forests outperformed the Boosted model.

model	***mse***	***var***	***error_rate***
random_forest	0.00252	0.0157	0.160
boosted	0.00303	0.0157	0.192

Conclusions

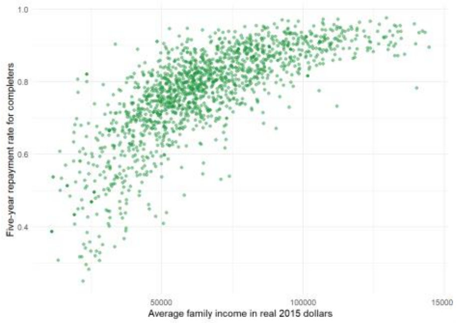
Our first question is answered. It is possible to predict federal loan repayment with some accuracy. However, our second question remains. What does that mean?



I think there are a few things we can draw from the results. First, the fact that decision trees were, at the end of the day, the best model is reassuring. Humans often think in decision trees. In fact, that is how College Scorecard is set up. We add filters that will sort the information as we seek some sort of match for the ideas in our head.

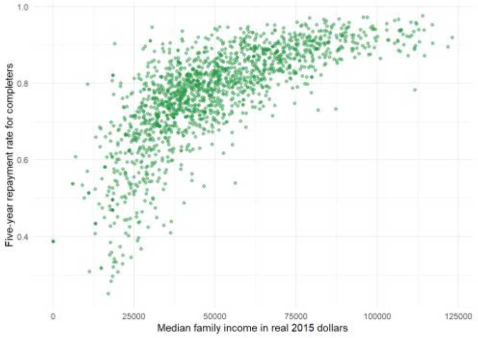
Based on the uses of the variables, the average family income, median family income and income of the student for a given institution is are three of the top predictors, and all follow the same general shape.

The reliance of the forest upon income is both enlightening and disappointing. Students often can't control the amount of money they have and attending a school with a high repayment rate may misleading since it is not base as much on other factors such as the major or diversity of the school.

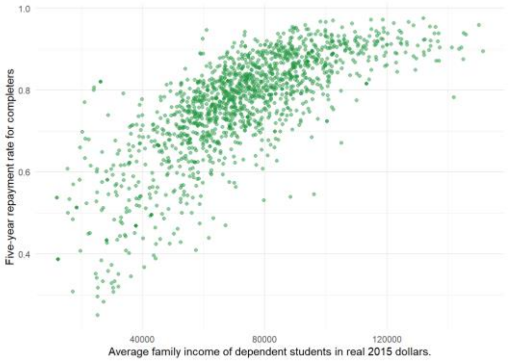


Future Questions

However, these variables are most likely strongly related, so a future project following the same question might consider removing some of these variables or thinking about collinearity more carefully.



Another consideration is the handling of missing data. There were missing values for the forests but since XGBoost handles that issue internally, it was easier to work with. It would be interesting to think more about the missing data, and even see if it is possible to predict when an institution will not report data.


Five-year repayment rate for completers
Average family income of dependent students in real 2015 dollars.