SAS® GLOBAL FORUM 2021

Paper #1196-2021

Investing In Education Intelligently

Sam Edison, Akhil Emani, Nithisha Katta, Rushya Puttam

Oklahoma State University.

ABSTRACT

Student loans are a critical resource used by many Americans to help pay for the goal of achieving higher education. As the tuition for post-secondary education continues to rise, the need for understanding student loans and the variables that impact individuals' abilities to pay them back is vital. In a recent survey, Country Financial found that only 9% of parents talk to their children about managing student loan debt (PR Newswire, 2019). Many borrowers are teenagers who have little formal financial education and even less experience in such financial matters. The College Scorecard is an online tool assisting students to evaluate higher education options. Created by the United States government, the tool helps address the disconnect experienced by future students and the high cost of advanced education. Using SAS, this study analyzes data compiled by the Scorecard to provide further insight into possible student outcomes based on institution-level data.

INTRODUCTION

Many high school students in the United States are excited to continue their education at post-secondary institutions, allowing them more freedom and the option to explore various subjects. This education frequently comes at a premium cost. Students must rely on hard-earned scholarships, grants, parent-funding, student loans, or a combination of these resources. U.S. News reported that 42 million, or one in six American adults, currently carries a federal student loan. Recent calculations for student loan debt in America are $1.6 trillion (U.S. News & World Report, 2019).

Students may not have the financial foresight, education, or awareness of possible career pathways to project the outcome of the cost of student loans versus possible earnings after graduation. This study is designed to evaluate post-secondary institution costs and common outcomes to increase the knowledge of current and future students looking to achieve their educational goals.

PROBLEM

The objective of this study is to identify likely outcomes for individuals seeking to invest in higher education. To achieve this, predictive and visual analytical tools are explored. In the incurrence of student loan debt, it is important to consider post-graduation income and the ratio of debt-to-income (DTI). The objective of this study is to identify likely outcomes (in the context of DTI) for individuals seeking to invest in higher education. Various factors that are likely to impact DTI such as degree acquired, U.S region, etc. are identified and their impact is explained. To achieve this, predictive and visual analytical tools are explored.

DATA

The dataset used for this study is compiled annually by the U.S. Department of Education,

currently available from academic years 1996-97 to 2018-19 (College Scorecard[1]). Each year is represented by an individual file containing aggregate data for each institution, including information on institutional characteristics, enrollments, student aid, costs, and student outcomes. The analysis provided in this report utilizes the 2014-15 academic year, containing over 7,000 rows (institutions) and nearly 2,000 variables. The dataset compiled for the 2014-15 academic year represented the most completed observations for the target variables. Many data elements within the College Scorecard are only available for federal financial aid recipients. These data are reported at the individual level to the National Student Loan Data System (NSLDS), which is used to distribute federal aid, and published at the aggregate institution level. Any elements containing aggregations with fewer than 30 students in the denominator are represented as "PrivacySuppressed" to ensure data privacy and representativeness.

DATA PREPARATION

Data Cleaning, Imputation and Reduction

Variables containing "NULL" tokens or "PrivacySuppressed" were replaced with standard SAS missing value representations. Institutions with 50% or more missing values were dropped from the dataset. Further, institutions that did not belong to one of the 50 U.S. states were removed. For instance, the universities in Puerto Rico and the Virgin Islands (University of the Virgin Islands-Albert A. Sheen) were removed.

Missing values were handled for both interval and categorical variables by imputing the median for interval variables and the mode for categorical variables. Since imputing with mean reduces the variance in the dataset, the median is chosen over mean for interval variables. Variables with missing values greater than 50% were removed from the dataset.

Data Derivation

New variables were derived from the existing variables. For better modeling, one hot encoding was applied for all the categorical variables. The universities were broken down into four distinct regions instead of 50 states. Thus, a region variable was created that buckets states into four regions: Northeast, Midwest, South, and West. The target variable DTI was also created using SAS code and will be addressed in the analysis section of the report.

Data Transformation

The interval variables in the dataset have values with varying ranges from small to large. The small values span the range (0, 1). The large values span the range (0, 100,000). To meaningfully compare all variables in the dataset. A series of transformations was applied to the large values to scale them down to be comparable with the small values.

ANALYSIS

TARGET VARIABLE

Median Debt defines the accumulated amount of federal loans received by all student borrowers. The median debt level is the aggregate of students' debt who complete a degree at an institution. Median Earnings is an important consideration for students and an indication of institution value. Students enroll in diverse programs of study; their earnings reflect the labor market's valuation of the education acquired in school. This measure provides the median earnings 6 to 10 years after enrolling in an institution while excluding students currently enrolled (*Full Data Documentation*, *36*).

[1] https://collegescorecard.ed.gov/data

Determining the return on investment of attending a post-secondary institution has certain dependencies that are assumed in this study, such as the student will graduate and join the workforce at least six years after beginning their education at an institution. Furthermore, the interest rate (5%) and term (10 years) used in the student loan payment calculation were generalized to avoid assumptions that rates will be in favor of the future student (*Federal Student Aid*). The target variable, *debt-to-income ratio* (DTI), is calculated by dividing the monthly payment by the monthly income as seen in the formula below:

*MonthlyPayment = p/t + (p * i) / 12*	*P* = Median Debt; *i* = interest rate (5%)
MonthlyEarnings = $E/12$	*E* = Median Earnings
DTI = MonthlyPayment/MonthlyEarnings	*t* = term (120 months)

Table 1. Calculation for DTI

VISUAL ANALYTICS

Target Variable Correlation

Figure-1 displays how individual institutions (represented by bubbles) can be categorized from the target variables: Earnings, Debt, and DTI. The median earnings are shown along the x-axis, while the y-axis displays the median debt. The DTI is represented as both the size of the bubble and color saturation, where large and red indicate a high DTI, as opposed to, blue and small being low DTI. Furthermore, the chart is broken down into four quadrants (indicated by the black lines). While most of the data appear to cluster near the center of the chart, there are instances of debt rising with earnings, demonstrating a moderate positive relationship (correlation: 0.3 to 0.5) between the Earnings and Debt variables.

Quadrant 2 (top left) describes institutions with low earnings but high debt. These institutions' students should take more caution and care when considering attending.

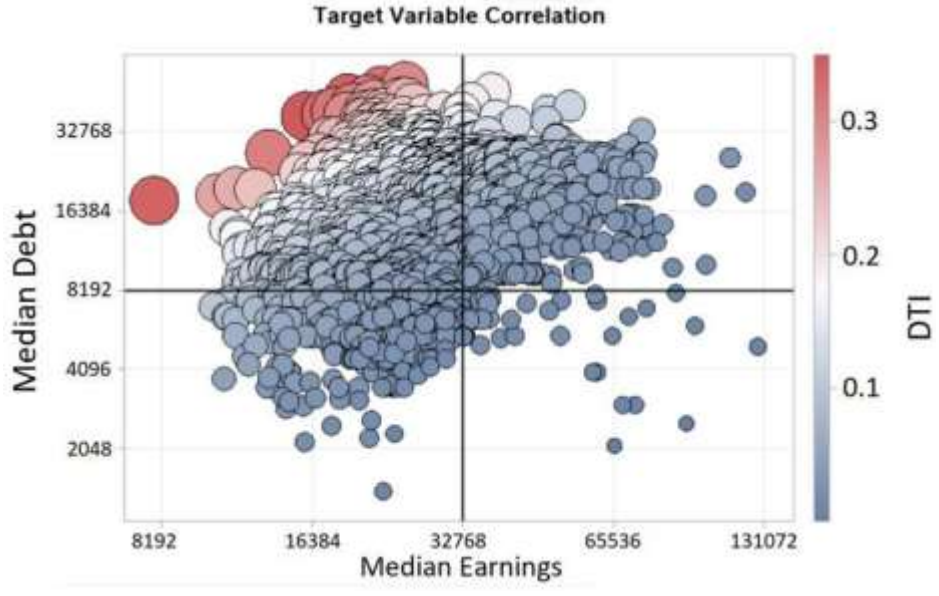

Figure 1. Target Variable Correlation

Quadrant 4 (bottom right) holds the best opportunities from a DTI perspective where each institution has high earnings and low debt. Unfortunately, this quadrant represents the fewest number of institutions in the dataset.

Median Earnings and Debt by Predominant Degree and Region

The predominant degree (See Figure-2 in the Appendix-A) identifies the type of award that an institution primarily allocates. The bars indicate the median earnings and the line represents the median debt. The desire is for the bar to exceed the line (lower DTI ratio). This occurs in the bar for Master's degrees, meaning that at institutions where Master's degrees are predominantly given, the DTI is lower. Certificates also rank well, as the line and bar are nearly connecting. Meanwhile, Associate's and Bachelor's display a deficit to be considered during decision-making.

Various Regions (See Figure-3 in the Appendix-A) were created from the State variable within the dataset (see appendix for detail). Using similar reasoning, the Northeast (N) and West (W) bars represent the best DTI across all regions. The Northeast shows the highest amount of debt but alternatively, the highest earnings. Meanwhile, the Midwest (M) and South (S) maintain higher debt with lower earnings.

MODELING

The variable selection node was used to identify the variation explained by each predictor variable in explaining the debt-to-income target. The variable selection yielded 30 variables remaining for modeling. To understand the effect of variable selection in the rejection model, two baseline regression models are run with and without variable selection. Mean Square Error (MSE) was chosen as the model comparison metric. The MSE measures the average of the squared distance between the predicted values and the actual values.

On the reduced dataset (30 variables), a series of models such as regression (with forward, backward and stepwise), random forest, a neural network with the default setting, a neural network with 2 hidden layers, a neural network with five neurons, and high-performance tree were built. All these models are run and compared using the model compare node with the criteria as Average Squared Error. Random Forest turned out to be the best model with an average squared error of 0.00020. To interpret the model results, the most important variables given by this model are passed into a regression model to understand the effect of the variables on the debt-to-income ratio.

The top three important variables resulted from the random forest model are: *Percent who transferred to a 4-year institution and withdrew within 4 years*, *Three-year repayment rate for first-generation students,* and *Percent of not-first-generation students who completed within 2 years at the original institution* (See Figure 4 in Appendix-A for the full list of important variables). The parameter estimates of these important variables resulting from the regression model show which factors influence the debt-to-income ratio positively and negatively. Here, the variables which have the negative estimate coefficients are favorable in decreasing the debt-to-income ratio which is the desired outcome. The variables that decrease the debt-to-income ratio are *Graduate Degree, Certificate Degree, Percentage of degrees awarded in Personal and Culinary Services,* and more (See Figure 5 in Appendix-A). The factors that increase the debt-to-income ratio are *Instate tuition fees*, *and Bachelor's degree in Communications Technologies/Technicians And Support Services.*

SUGGESTIONS FOR FUTURE STUDIES

This study focused solely on post-graduation outcomes for debt and earnings for graduating students. Further assessment of an institution's acceptance and completion rates used in conjunction with the student's test scores and high-school GPA could suggest a future

student's likelihood to succeed, regardless of cost. Furthermore, the College Scorecard has become a widely used tool that is still being expanded. A secondary dataset, *field of study*, began forming in 2019 to include disaggregated data elements describing post-graduation earnings and cumulative loan debt of graduates by field of study and degree earned. This disaggregated data would provide more granular insights into the student outcomes for each field of study and degree types at a specific university, as opposed to, the aggregated version of the Scorecard used in this study. For example, a student could compare an Engineering major to a Computer Science major at a given institution.

CONCLUSION

While the aggregated dataset does not provide granular detail for modeling, many results were identified that can further educate individuals on the subject of student loans. It was seen from the descriptive analyses that institutions predominantly offering Master's Degrees and Certificates represented a lower DTI than institutions predominantly offering Associate's and Bachelor's degrees. Regions, such as the Midwest and South, demonstrate higher DTI's due to lower earnings than the Northeast and higher debt than the West.

The champion model reiterated that future students should seek Master's Degrees and Certificates to lower DTI. Additionally, pursuing a degree in Personal or Culinary services offers a lower expected DTI. The factors increasing DTI were partly due to institutions with high In-state tuition and fees, as well as, institutions predominantly offering bachelor's degrees.

REFERENCES

Financial, COUNTRY. "Parents Just Don't Understand (Finances): 3 in 5 Americans Rely on Uncertain Parents for Financial Guidance." *PR Newswire: News Distribution, Targeting and Monitoring*, 18 Sept. 2019, www.prnewswire.com/news-releases/parents-just-dont-understand-finances-3-in-5-americans-rely-on-uncertain-parents-for-financial-guidance-300920509.html.

Hayes, Adam. "Multicollinearity." *Investopedia*, Investopedia, 22 Oct. 2020, www.investopedia.com/terms/m/multicollinearity.asp.

***Federal Student Aid**, studentaid.gov/understand-aid/types/loans/interest-rates.*

***Federal Student Aid**, studentaid.gov/manage-loans/repayment/plans.*

"Student Debt Explained: Breaking Down the $1.6T in Loans." *U.S. News & World Report*, www.usnews.com/news/elections/articles/2019-11-01/student-debt-explained-breaking-down-the-16t-in-loans.

College Scorecard, Office of Planning, Evaluation and Policy Development (OPEPD), collegescorecard.ed.gov/.

Full Documentation, Office of Planning, Evaluation and Policy Development (OPEPD), https://collegescorecard.ed.gov/assets/FieldOfStudyDataDocumentation.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Samuel Edison
Samuel.edison@okstate.edu

Akhil Emani
aemani@okstate.edu

Nithisha Katta
nkatta@okstate.edu

Rushya Puttam
rputtam@okstate.edu

APPENDIX-A: MODELING

The below table shows the mean square error of regression models run with and without variable selection before proceeding to run a series of models.

SL. No	Model	Mean Square Error
1	Regression with variable selection	0.0149
2	Regression without variable selection	0.0370

Table 2. MSE With and Without Variable Selection

The below table shows the average squared error for the series of models run.

SL. No	Model	Average Squared Error: Validation
1	Random Forest	0.00020
2	Neural Network with 5 neurons	0.00026
3	Neural Network	0.00029
4	Decision Tree	0.00031
5	Neural Network with two hidden layers	0.00031
6	Regression: Stepwise selection	0.00039
7	Regression: Forward selection	0.00039
8	Regression: Backward selection	0.00039

Table 3. Model Performance Metrics

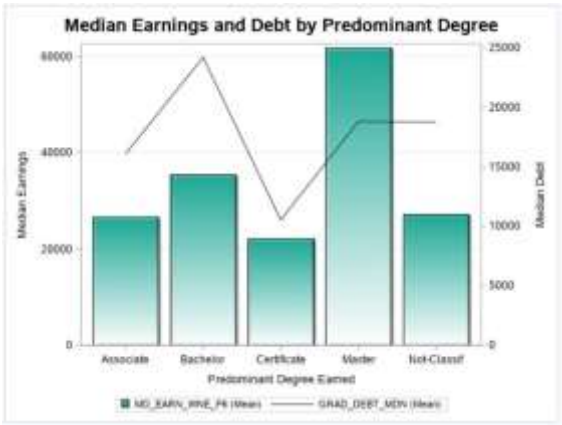

Figure 2. Predominant Degree

Figure 3. Target Variable Correlation

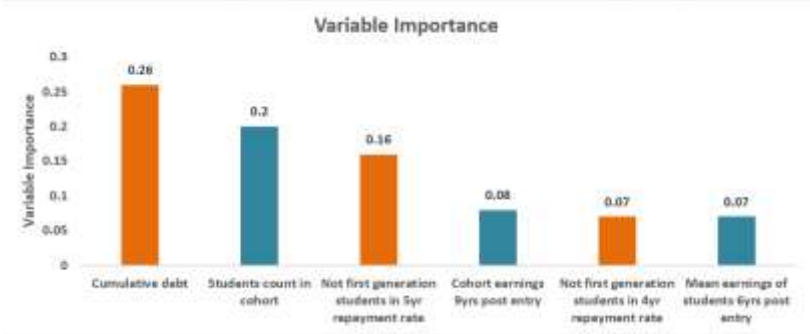

Figure 4. Variable Importance

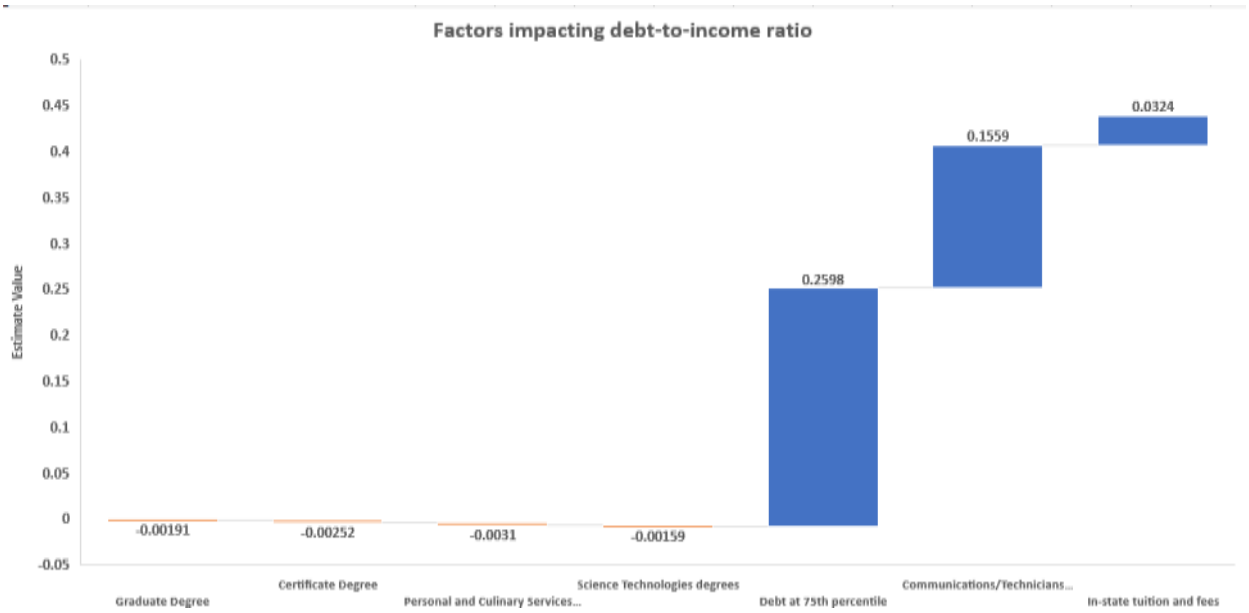

Figure 5. Factors impacting debt-to-income ratio

APPENDIX-B: DATA REDUCTION

Variable Clustering is used to address the multicollinearity assumption, whereby predictor variables should not be collinear with each other. This node helps to reduce the data as well as to remove the collinear variables. The settings used here are determining clusters by hierarchical clustering method, not confining any value to maximum clusters. Two stage clustering is set to "yes" as the dataset is large, with more than 200 variables. The Cluster split criterion is set to Second Max Eigen Value > 1. Typically, this node is for numeric variables, however, class variables can be used after changing to dummy variables. Variable clustering is performed to identify the variables which fall into similar clusters. Global clusters are formed with the following formula:

$$\text{Number of clusters} = \text{INT}((\text{number of variables} / 100) + 2)$$

After the data imputation node, there are 714 variables, and substituting this in the above formula gives the number of clusters as 9.

$$\text{Number of Clusters} = \text{INT}((714/100) + 2) = \text{INT}(7.14+2) = 9$$

Variable clustering is then performed on the global clusters. The 9 global clusters comprise of 43 clusters. Variables from each cluster are selected using 1- R2 ratio. The lower the ratio, the higher the chance of selection.