Stat 159 Final Project - Providing Credit to Students

Team W.O.S: Yoon Jung Rho, Young Hoon Kim, Ellen Hwang, Joseph Simonian

December 6, 2016

Abstract

Our client profile is a credit institution that provides financial aid to students. The managers are interested in expanding their customer base but they would also like that most of the loans be payed back. Our team's purpose is to perform exploratory data analysis and create predictive models to find which schools and what kind schools credit institutions should provide credit.

1 Introduction

Our client, a credit institution that provides financial aid to students, want to expand their customer base but would also like that most of the loans be payed back. Our role as the analyst is to use the publicly available [College Scoreboard Datasets](https://collegescorecard.ed.gov/data/) to figure out what features of a college make it more reliable for credit. Our team will be using the 3 year repayment rate as am indicator of the school's overall reliability rate. Using exploratory data analysis we will examine the relationships between repayment rate and other features of the school. In our analysis, we will use ridge, lasso, partial least squares, and principle component regression to find significant features that influence a school's overall repayment rate.

2 Data

The data from College Scorecard provide insights into the performance of schools eligible to receive federal financial aid, and offer a look at the outcomes of students at those schools. The Data that appear on the College Scorecard provides data on student completion, debt and repayment, earnings, and more. The files include data from 1996 through 2016 for all undergraduate degree-granting institutions of higher education. This data was last updated on September 13th, 2016. The data is available at: https://collegescorecard.ed.gov/data/

For our project, besides the main data, our team also used featured downloads provided by College Scorecard. These data downloads provide quick access to some of the data in which users may be most interested, including a file that offers the most current data for each element. Among variety of data, we used Post-School earnings data set (https://ed-public-download.apps.cloud.gov/downloads/Most-Recent-Cohorts-Treasury-Elements.csv) to narrow down the analytical component.

There is also a documentation that provides more on how to use the data, including: Sources of the data, The construction of metrics, and Data considerations and limitations available at: https://collegescorecard.ed.gov/data/documentation/

3 Methods

The data that we use for the analysis contains 7703 universitites under 1743 different categories. In order to analyze the data more efficiently, our team decided to remove all the columns that doesn't have sufficient data; we considered the columns as insufficient if 99 percent or more of the column values are NA. After this process, we decreased the number of the columns to about 500. We picked 3yr repayment rate as our response variable and after the first round of cleaning, the team went through all the columns by each category and picked out columns manually that are related to client's profile and response variable. List of the variables can be found under data_cleaning_script.R file. After we obtained specific list of column names, we converted all the

NULL values and PrivacySuppressed to NA. After removing Null and PrivacySuppressed values, we removed all columns with at at most 50

After ols, we removed variables with greater than 0.05 pvalue produced clean data as csv file under clean_data folder.

4 Results

4.1 Correlation and Regession Results to find Relevant Variables

Below is the table that contains predictors with absolute correlation value above 0.5. We decided that predictors with absolute correlation values above 0.5 would have influence that is significant enough in explaining repayment rate.

Table 1: Predictors with high correlation

	Variable	Correlation.Value
1	INC_PCT_H1	0.81994
2	INC_PCT_H2	0.76330
3	INC_PCT_M2	0.76211
4	C200_4_POOLED_SUPP	0.71878
5	MN_EARN_WNE_INC1_P6	0.65363
6	C150_4_POOLED_SUPP	0.64861
7	PCT25_EARN_WNE_P6	0.64779
8	PCT10_EARN_WNE_P6	0.64555
9	GT_25K_P6	0.63666
10	C150_4	0.61578
11	MD_EARN_WNE_P6	0.60904
12	NPT4_75UP_PUB	0.60720
13	MN_EARN_WNE_P6	0.58736
14	TUITIONFEE_OUT	0.58227
15	PCT75_EARN_WNE_P6	0.57419
16	C150_4_WHITE	0.56702
17	NPT4_PUB	0.56050
18	C100_4	0.55703
19	PAR_ED_PCT_1STGEN	-0.71090
20	CDR3	-0.75043
21	INC_PCT_LO	-0.83469

Another measure that we employed in deciding predictors that explain our response variable well was running regression. We believe that the factors that act as good predictors of repayment rate on federal loans will also act as good predictors of repayment rates on loans given by our client. So, we create regression models that predict repayment rate of federal loans based on other public features from the scorecard data.

In order to select the predictors and their coefficients that best account for the repayment rate, we ran Ordinary Least Squares, Lasso, Ridge, Principal Components, and Partial Least Squares regressions. From those five regressions, we compared the test mean-squared-error rates (MSE) of each regression. Below is the MSE table for all the regressions.

Table 2: MSE of regressions

	OLS	Lasso	Ridge	PCR	PLSR
1	0.04325	0.01519	0.01504	0.01507	0.01506

It is not difficult to observe that ridge regression has the lowest MSE value out of all regressions. Below is the regression predictors and corresponding coefficients from ridge regression.

Table 3: Ridge Coefficients

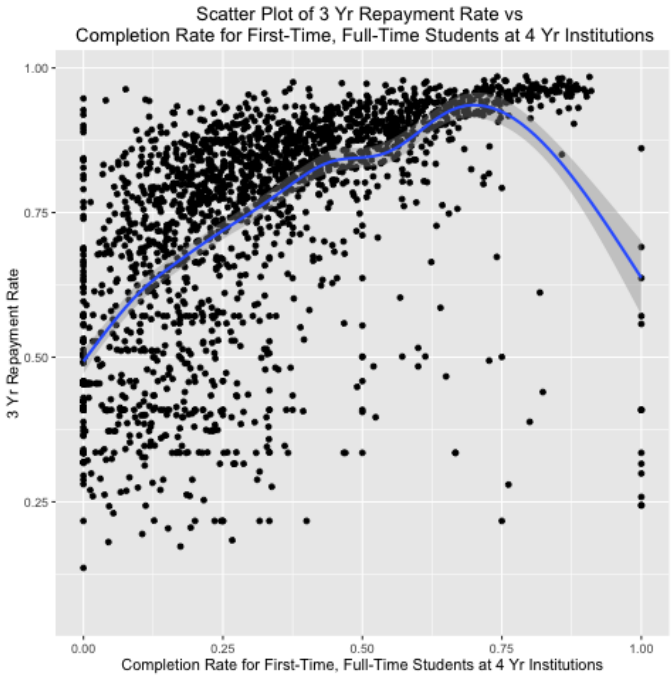
	Coefficients
INC_PCT_M2	0.0218773
C150_4_POOLED_SUPP	0.0128558
PCT10_EARN_WNE_P6	0.0000002
GT_25K_P6	0.0074061
C150_4	0.0110158
NPT4_75UP_PUB	0.0000009
TUITIONFEE_OUT	0.0000001

According to the table above, our client should pay close attention to:
1. Percentages of aided students with family incomes between $48,001-75{,}000$ in nominal dollars
2. 150 percent completion rate for less-than-four-year institutions, pooled in two-year rolling averages and suppressed for small n size
3. 10th percentile of earnings of students working and not enrolled 6 years after entry
4. Share of students earning over 25,000 dollars per year (threshold earnings) 6 years after entry
5. Completion rate for first-time, full-time students at four-year institutions (150 percent of expected time to completion/6 years)
6. Average net price for family with more than 75,000-dollar income to pay for the education for public institution
7. Out-of-state tuition and fees

4.2 Exploratory Data Analysis

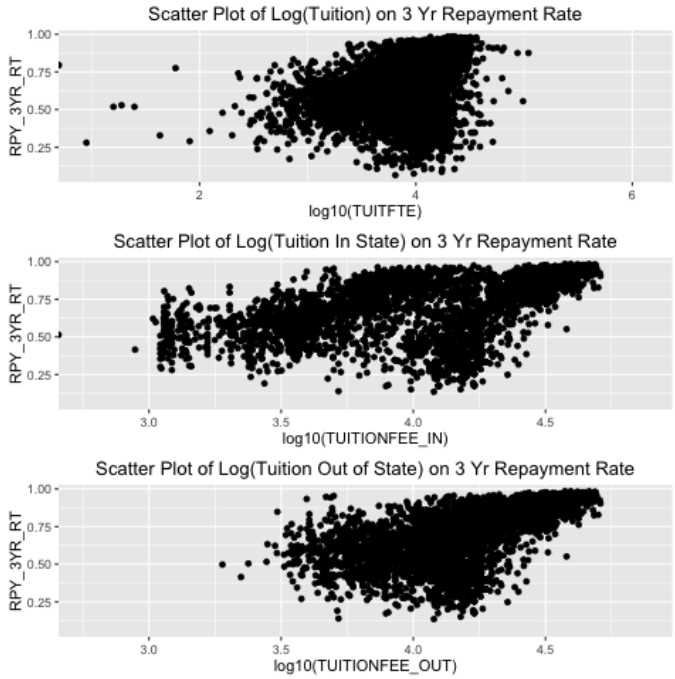
After using correlation and regression to select features related to repayment, we were able to perform some exploratory data analysis. We have listed some important visualizations on an online Shiny interface and below.

Figure 1: Scatter Plot of Completion Rate vs. 3 Year Repayment Rate



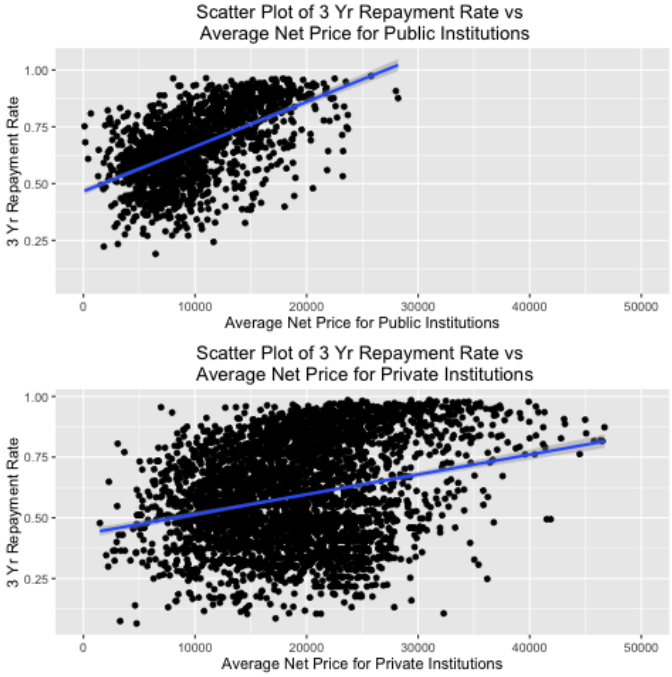
From the graph, we can see that schools with repayment rates below 75% tend to have completion rates below 50%. There is also a very steep trend for completion rate for repayment rates above 75%.

Figure 2: Scatter Plot of Different Tuition Types on 3 Year Repayment Rate



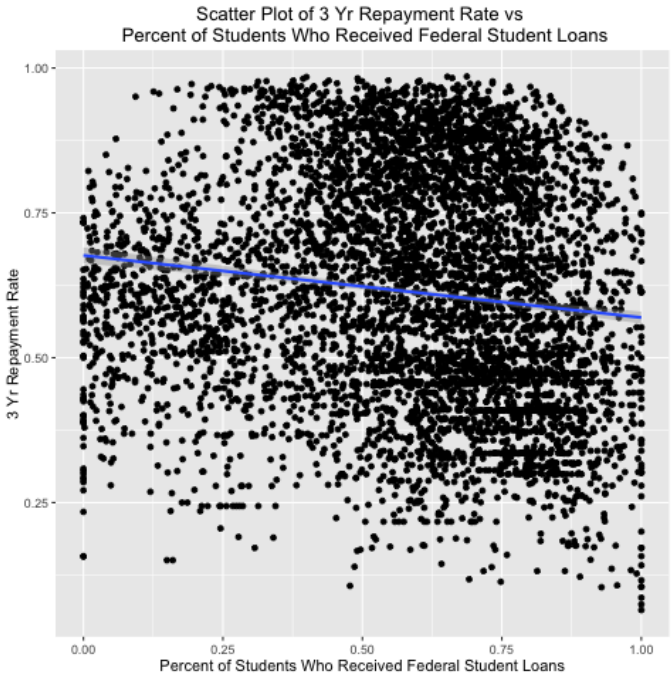
This figure displays three different scatterplots of Tuition fee vs Repayment Rate. In the data set, there are three different information of Tuition fee under cost category. First graph is a Net tuition revenue per full-time equivalent student vs repayment rate. Second graph is a In-stat tuition and fee vs repayment rate and last one is out-of-state tuition and fees vs repayment rate. There is highest correlation between out-of-state vs repayment rate.

Figure 3: Scatter Plot of 3 Year Repayment Rates on Type of Institution



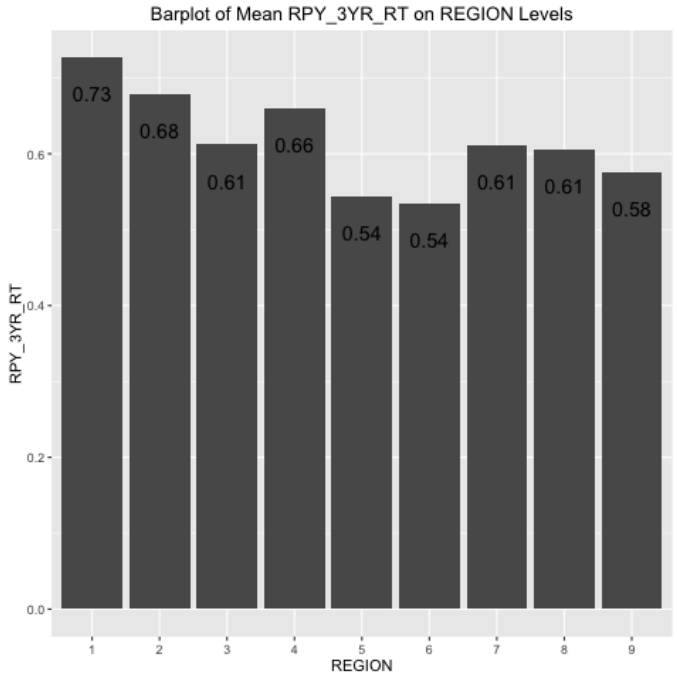
The figure above is a scatterplot of 3 Year Repayment Rate vs average net price for public institutions and below, it has a scatterplot of 3 Year Repayment Rate vs average net price for private instutitions. It shows higher slope for public institution but for both graph, it shows weak relationship between the average net price and repayment rate.

Figure 4: Scatter Plot of 3 Year Repayment Rates vs Percent of Students Who Received Federal Student Loans



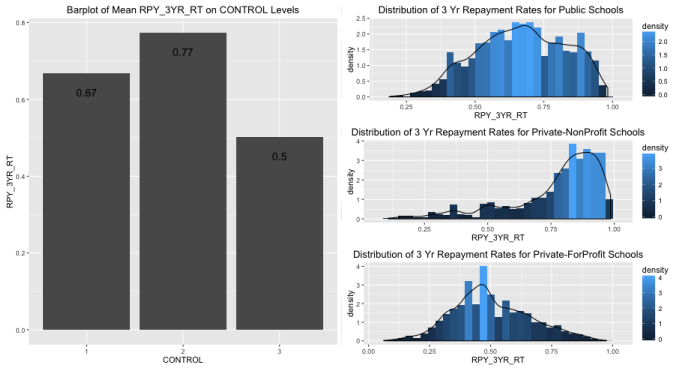
The figure above is a scatterplot of 3 Year Repayment Rate vs percent of students who received federal student loans. By looking the graph, it shows that there is weak relationship between percent of students who recieved federal student loans and 3yr repayment rate.

Figure 5: Barplot of Mean 3 Year Repayment Rate on REGION Levels



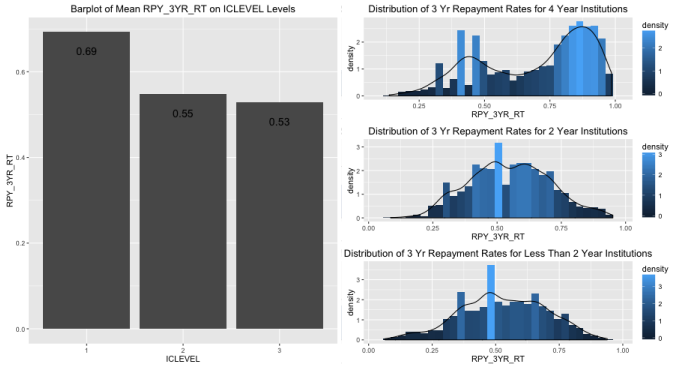
This figure displays a barplot of average value of 3 Year Repayment Rate for each region. 3 Year Repayment Rate is defined as a fraction of repayment cohort who are not in default, and with loan balances that have declined three years since entering repayment, excluding enrolled and military deferment from calculation. (rolling averages) and each university is divided into 9 regions: 0 U.S. Service Schools, 1New England (CT, ME, MA, NH, RI, VT), 2 Mid East (DE, DC, MD, NJ, NY, PA), 3 Great Lakes (IL, IN, MI, OH, WI),4 Plains (IA, KS, MN, MO, NE, ND, SD), 5 Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV), 6 Southwest (AZ, NM, OK, TX), 7 Rocky Mountains (CO, ID, MT, UT, WY), 8 Far West (AK, CA, HI, NV, OR, WA), 9 Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI). Our graph is not showing region 0 because there is one university that is under region 0 and it did not have a value for 3 Year Repayment Rate On barplot it shows that region 1 and 2 have highest repayment rate(0.73 and 0.68, respectively) and region 5 and 6 have lowest repayment rate (0.54)

Figure 6: Barplot of Mean 3 Year Repayment Rate on Control Levels and Distributions of 3 Year Repayment Rate for Control Level



This figure displays a barplot that shows average value of 3 Year Repayment Rate for each school type. Under Control column, each school is divided into three categories: 1 = Public, 2 = Private nonprofit, 3 = Private for-profit. It shows that Private for nonprofit universities has highest average repayment rate with 0.77. On the right side, histogram helps for the better understanding of barplot. The visualizations on the right shows density graph and histogram of each institution's repayment rate.

Figure 7: Barplot of Mean 3 Year Repayment Rate on ICLEVEL Levels and Distributions of 3 Year Repayment Rate for each ICLEVEL Level



This figure is a barplot that shows average value of 3 Year Repayment Rate for each level of institution. Uncer ICLEVEL, each school is divided into three categories: 1 = 4-year,2 = 2-year, 3 = Less-than-2-year. It shows that 4-year institution has the highest average repayment rate with 0.69. On the right side of the barplot, there is a histogram/density plot of each institution's repayment rate for detailed analysis.

5 Analysis

In order to provide advice on future strategies to our client, we thought it proper to conduct *time series analysis* on all the schools with repayment rate of 3 years available to understand their future repayment rates. Unfortunately, not many years of data files were available as most of earlier data files did not have any data for repayment rates; however, we were able to observe repayment rate of 3 years of data starting from 2009. Notice that since there is only 6 years of data available for each school, the time series analysis on each school is mostly a way of understanding the trend of each school within those 6 years.

With the time series forecast of three years on each school, we were able to pick 'Top 100 Schools' according to their mean future 3-year repayment rates. Below is the table of the top ten schools with the best projected repayment rates.

Table 4: Top 10 Schools

	INSTNM	row_mean
443331	West Coast University-Los Angeles	1.0000
167996	Stonehill College	0.9936
217165	Bryant University	0.9924
213251	Juniata College	0.9909
177719	Barnes-Jewish College Goldfarb School of Nursing	0.9892
211440	Carnegie Mellon University	0.9856
239716	Saint Norbert College	0.9824
152080	University of Notre Dame	0.9792
181428	University of Nebraska Medical Center	0.9788
216524	Ursinus College	0.9785

Notice the first school on the list has the average forecasted repayment of 1. This was fixed from its original value of 1.017357 since repayment rate cannot exceed 1. Such anomaly can be explained by the insufficient data availability for each school. According to the time series analysis, the cutoff for the projected repayment rate top 100 schools was 0.9491918, and the school name is College of the Holy Cross.

6 Conclusions

In this report, we analyzed how repayment rates on federal student loans varied as a result of several factors. We analyzed how federal student loan repayment rates vary between public and private schools, between regions, and between colleges with differing tuition costs and completion rates. In addition, we used regression and time series analysis to predict which colleges will likely have high private loan repayment rates, and which colleges have an upward trend in repayment rates - information valuable to a customer such as ours.

Our regressions indicated that our client should pay closest attention to student completion rates, as these were the best predictor of repayment on federal loans for both four-year and less-than-four-year institutions. In addition, our regression analysis showed certain features of post-graduate income to be a good predictor of repayment. Specifically, while average postgraduate income was a poor predictor of repayment rates, we found that the share of students earning over $25000 per year was a good predictor of repayment. From this information, it is reasonable to conclude that students earning over a certain cost-of-living threshold will be much more likely to pay back their loans than students making less, but that loan repayment rates are somewhat steady above a certain threshold. While we do not have the student-by-student data to verify this, we can recommend that our client focus not on average postgraduate income but on the percentage of graduates making above a L̈iving Wage.̈

Finally, we performed time series analysis to determine which schools are likely to have the highest future repayment rates, based on their past and current rates over six years. We created a list of top 100 colleges by predicted future repayment rates - our client can use this list to identify schools that are likely to well better in the future, in order to shift their efforts to those schools.