Stat 159 Final Project - Providing Credit to Students

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

December 5, 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](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. (python?)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 95 percent or more of the column values are NA. After this process, we decreased the number of the columns to about 500. After the first round of cleaning, the team went through all the columns by each category that are related to client's profile. 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 Privacy-Suppressed to NA. After removing Null and PrivacySuppressed values, we performed correlation matrices to see which predictors have high correlation between 3yr repayment rate. Since all of the

repayment rate has high correlation with 3yr repayment rate, we took out rest of the repayment rate columns and picked variables that has higher absolute value of 0.5 correlation coefficient. With specified variables, we ran basic ols regression to see which variables to and which to clean out(we are not using them tho).

Table 1: MSE of regressions

	OLS	Lasso	Ridge	PCR	PLSR
1	0.04325	0.01519	0.01504	0.01507	0.01506

```
ridge_coef <- xtable(as.data.frame(rpy3yr_rr_coef[-1,]), caption = 'Ridge Coefficients', digits =
print(ridge_coef, comment = FALSE, table.placement = 'H', type = 'latex', caption.placement = 'top
```

Table 2: Ridge Coefficients

	rpy3yr_rr_coef[-1,]
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

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

WRITE SOMETHING ABOUT USING REGRESSIONS TO SELECT VARIABLES

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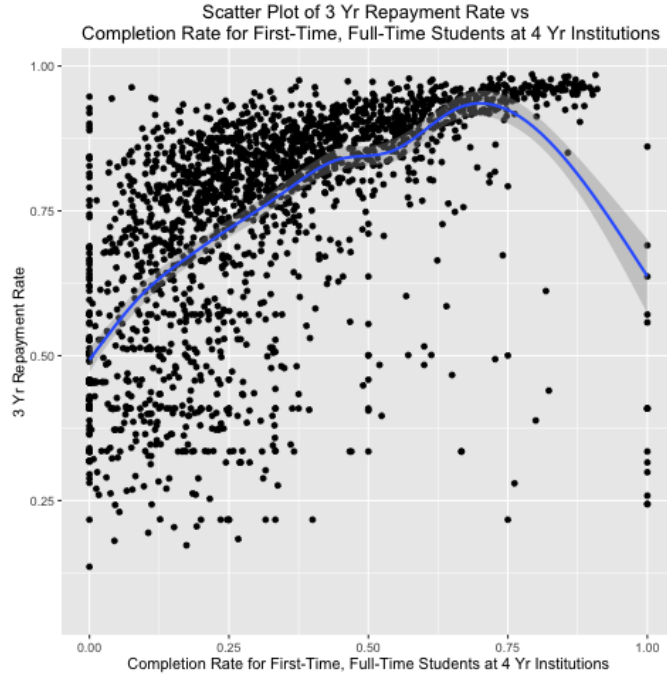
From the graph, we can see that schools with repayment rates below 75% tend to have completion rates below 50%. There is also a very steep trend for completion rate for repayment rates above 75%.

Figure 2: Scatter Plot of Different Tuition Types on 3 Year Repayment Rate
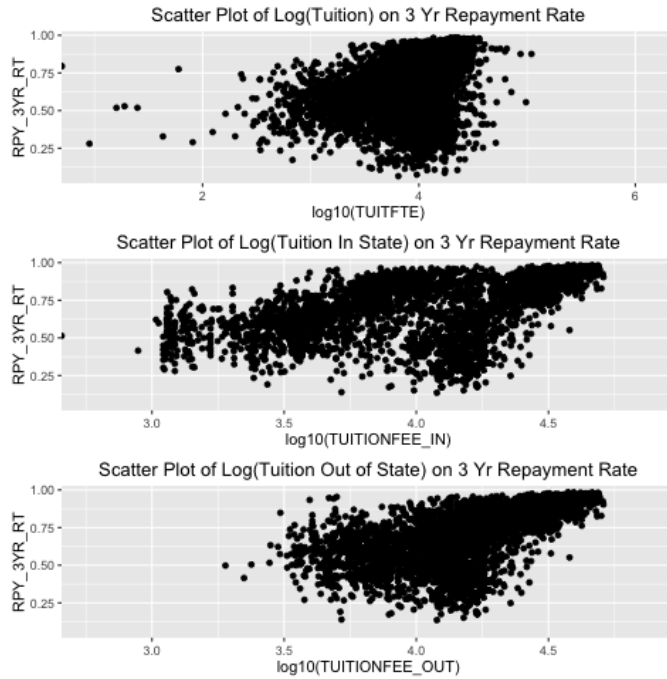
This figure displays three different scatterplots of Tuition fee vs Repayment Rate. In the data set, there are three different information of Tuition fee under cost category. First graph is a Net tuition revenue per full-time equivalent student vs repayment rate. Second graph is a In-stat tuition and 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.

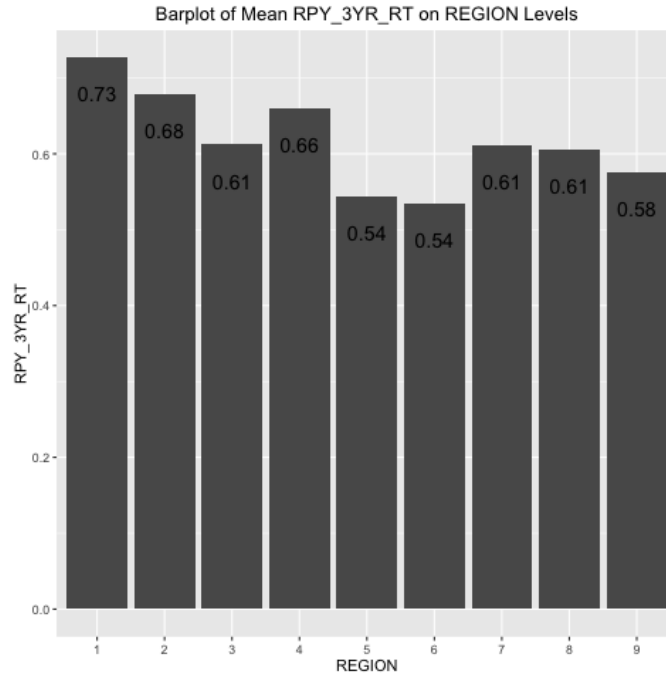Figure 3: Scatter Plot of 3 Year Repayment Rates on Type of Institution

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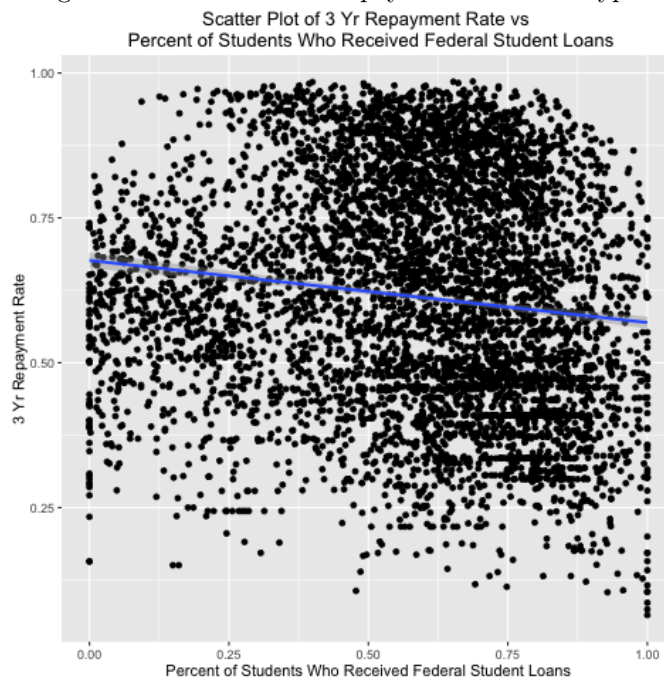
The figure aboce is a scatterplot of 3 Year Repayment Rate vs percent of students who received federal student loans. By looking the graph, it shows that there is weak relationship between percent of students who recieved federal student loans and 3yr repayment rate.

4

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

Barplot of Mean RPY_3YR_RT 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
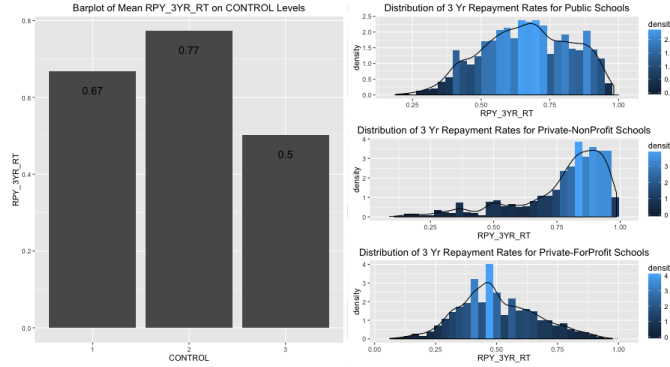
This figure displays a barplot that shows average value of 3 Year Repayment Rate for each school type. Under Control column, each school is divided into three categories: 1 = Public, 2 = Private nonprofit, 3 = Private for-profit. It shows that Private for nonprofit universities has highest average repayment 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
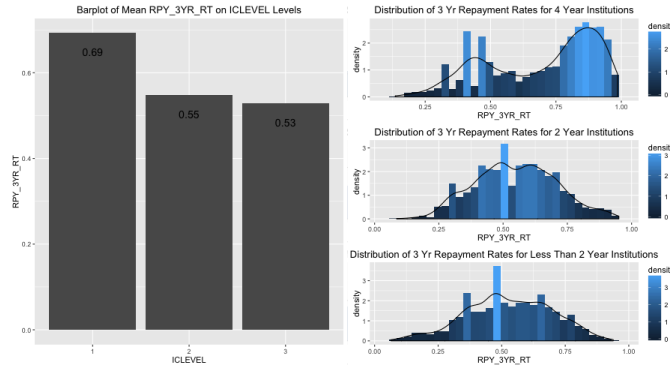
This figure is a barplot that shows average value of 3 Year Repayment Rate for each level of institution. Uncer ICLEVEL, each school is divided into three categories: 1 = 4-year,2 = 2-year, 3 = Less-than-2-year. It shows that 4-year institution has the highest average repayment rate with 0.69. On the right side of the barplot, there is a histogram/density plot of each institution's repayment rate for detailed analysis.

5   Analysis

Since repayment rate for each school is already available in our data set, there did not exist any need for prediction. However, since we are to provide advice to our client, a credit institution that provides financial aid to students, 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 rate. Unfortunately, not many years of data was available; 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 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 3: Top 10 Schools

	INSTNM	row_mean
443331	West Coast University-Los Angeles	1.02
167996	Stonehill College	0.99
217165	Bryant University	0.99
213251	Juniata College	0.99
177719	Barnes-Jewish College Goldfarb School of Nursing	0.99
211440	Carnegie Mellon University	0.99
239716	Saint Norbert College	0.98
152080	University of Notre Dame	0.98
181428	University of Nebraska Medical Center	0.98
216524	Ursinus College	0.98

Notice the first school on the list has the average forecasted repayment rate above 1, which is impossible. This can be explained by the insufficient data availability

6 Conclusions