US Student Debt

Ain Sanchez

2023-12-02

Introduction

Student debt in the United States has become a major problem in the last years, since many higher education institutions started increasing their costs more than inflation, while salaries in the job market remained stagnant. According to an article published by The Economist in early 2020, before the outbreak of the pandemic caused by COVID-19, the accrued student debt in the United States amounted nearly 1.5 trillon USD (The Economist, 2020). The World Bank also shows that the US gross-domestic product (GDP) reached 21.38 trillon USD in 2019, so student debt represented 7% of the US GDP in 2019 (The World Bank, 2023). This problem could have been severely aggravated during the pandemic, but the US Federal Government put student-debt repayment on hold until September 2023, which certainly gave a cushion for many student-debt holders in the past few years. The goal of this project is to predict the repayment behavior 7 years after entering repayment of student debt obtained from attending a higher education institution in the US.

Section 1: Methods and Analysis

To have a deeper look into the US student debt, the Department of Education periodically publishes datasets on the National Student Loan Data System (NSLDS) such as College Scorecards, which are intended to inform students about what types of colleges may be a good fit for them. These datasets contain information aggregated at two levels: institution and field of study. For the purpose of this project, only the data at the institution level will be used in order to capture the factors within a higher education institution that may contribute to student-debt repayment. Also, it is important to highlight that this project uses two publicly available datasets published on Kaggle, related to the NSLDS and College Scorecard, which have been previously cleaned and updated in March 2016 (Department of Education, 2016).

The NSLDS dataset provided by Kaggle contains 7,703 observations and 1,196 variables. Since the data is aggregated at the institution-level, the 7,703 observations refer to higher education institutions and their available branches in the US. Moreover, the dictionary of variables and the data documentation from the Department of Education's website were useful to understand the codification behind each variable name on the dataset (Department of Education, 2023). Based on this review, it could be determined that variables are grouped on different components related to: aid, completion, repayment, root/identification, school/name and student.

Within the repayment group of variables, 4 variables reflect the fraction of individuals who are not in default and have loan balances that have declined within 1, 3, 5, or 7 years after entering repayment. An article from the Best Colleges website, states that federal student loans can take 5 to 20 years to be repaid, so taking the cohort of individuals who entered repayment 7 years ago will better capture the repayment behavior as a predicted variable (Welding, L, 2023). Considering that 99 variables from the repayment group are related to the cohorts of 1, 3, and 5 years since entering repayment, they are eliminated from our dataset.

Since the completion component also included 982 variables with cohorts different from the the predicted variable's cohort, they were eliminated from the dataset. It is important to highlight that the completion

group of variables was completely eliminated from the dataset due to other cohorts' desaggregations. The other groups of variables that refer to the cohort of 7 years since entering repayment, or have some disaggregations related to the target cohort, contain 115 variables which are taken from the NSLDS dataset. Finally, 113 variables from the Scorecard dataset are also added to the NSLDS dataset, which results into a Student-Debt dataset of 7,703 observations and 228 variables. The following sections will describe the processes employed to prepare the data, run the model and discuss its results.

Data-Cleaning

The resulting dataset undergoes a data-cleaning analysis in terms of completeness, consistency and clarity.

Due to the large number of inconsistent variables in our dataset with respect to their variable type, the consistency analysis is first implemented. Based on the consistency analysis and the dataset documentation, it is evident that most variables are numeric but the NSLDS replaces some numeric values with a "PrivacySuppressed" label, whenever the system detects data that do not meet reporting standards. Thus, transforming these variables from character to numeric make the values labeled as "PrivacySuppressed" to be treated as missing values by coercion. The numeric variables that are indeed categorical are also transformed back to character, yet the completeness analysis becomes essential to determine the impact of missing data on our dataset.

In order to proceed with the completeness analysis, it is important to filter the observations which have empty values in the predicted variable (rate of repayment for the cohort of 7 years since entering repayment). Moreover, the Scorecard dataset introduces a variable to determine if an institution is currently operating or not, so having operating institutions is highly desired in order to obtain robust results. Upon filtering the empty values of our predicted variable and the non-operating institutions, the number of observations is reduced to 5,076, which represent operating institutions that actually reported data for our target cohort.

To carry out a comprehensive completeness analysis, it is crucial to focus on variables which have a very small percentage of missing values on the filtered dataset. While most variables on the dataset are numeric, the categorical variables tend to be more complete because they describe the institutions' demographics. As a result, numeric variables will be analyzed on the assumption that they follow a normal distribution, so a treshold of less than 0.5% (or +/- 3 standard deviations from the mean) of missing values is used as an acceptance parameter. The completeness analysis shows that only 23 out of the 228 variables (around 10% of the variables in the dataset), comply with the acceptance parameter, so these variables will be prioritized to avoid variable imputation.

The dataset shows clarity with 23 variables distributed into 5 NSLDS institutional identification variables (index, UNITID, OPEID, OPEID6, and INSTNM). Also, 10 numeric variables related to repayment rate (RPY_7YR_RT), median debt when entering repayment (DEBT_MDN), number of debt-holders (DEBT_N), number of cumulative debt-holders (CUML_DEBT_N), number of students who report family income (INC_N), number of individuals in the dependency status cohort (DEP_STAT_N), number of students with FAFSA applicants (APPL_SCH_N), average age of entry (AGE_ENTRY), average family income (FAMINC), and median family income (MD_FAMINC). Finally, 8 character variables that describe the institution such as city (CITY), state (STABBR), URL (INSTURL), net price calculator URL (NPCURL), status of cash monitoring by the Departmnet of Education (HCM2), preferred degree type (PREDDEG), financial source (CONTROL), and operating status (CURROPER).

Data-Exploring

Data-exploring is useful to briefly showcase the content of each variable in the new dataset with 5,076 observations and 23 variables. Considering the data type of each variable, either summary statistics or frequency tables will be used. In the case of numeric variables, summary statistics show values that include minimum, 1st quartile, mean, median, third quartile, maximum, and number of missing values. In the case of categorical variables, frequency tables come handy to display the distribution of the different value categories

within each variable. The following summary statistics were taken from the 10 numeric variables that were listed in the previous section:

```
# Apply summary statistics to numeric variables
for (i in 6:15){
  print(names(studentDebtFinal)[i])
  print(summary(studentDebtFinal[,i]))
}
```

```
## [1] "RPY_7YR_RT"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1149  0.4509  0.5741  0.5850  0.7210  0.9688
## [1] "DEBT_MDN"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1750    7815   10000   11868   15500   37500      25
## [1] "DEBT_N"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      10     437    1270    6180    3936  234259      16
## [1] "CUML_DEBT_N"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      10     437    1270    6180    3936  234259      16
## [1] "INC_N"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     14.0    488.2   1421.5   5356.1   4308.5 140665.0        2
## [1] "DEP_STAT_N"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     14.0    488.2   1421.5   5356.1   4308.5 140665.0        2
## [1] "APPL_SCH_N"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##     14.0    488.2   1421.5   5356.1   4308.5 140665.0        2
## [1] "AGE_ENTRY"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17.65   23.02   25.72   25.79   28.31   42.56       2
## [1] "FAMINC"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3373   23324   32046   40122   51631  142280       2
## [1] "MD_FAMINC"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   15574   21986   29327   36966  122446       2
```

The summary statistics reveal some interesting patterns among the numeric variables. Based on the results, it is evident that the variable CUML_DEBT_N is equivalent to the variable DEBT_N, while the variables DEP_STAT_N and APPL_SCH_N are equivalent to the variable INC_N. Thus, CUML_DEBT_N, DEP_STAT_N and APPL_SCH_N can be eliminated from the final dataset, because they do not add additional information to train the predictive model. On the other hand, some key variables show that the median value for average age of entry (AGE_ENTRY) at the institutional level is 25.72, while the median value for the average family income (FAMINC) at the institutional level is 32,046 USD. The following tables show the behavior of the categorical variables.

```
# Construct frequency tables to non-numeric variables, when applicable

# Checking at the variable CITY
sum(duplicated(studentDebtFinal$CITY) == F)
```

```
## [1] 2003
```

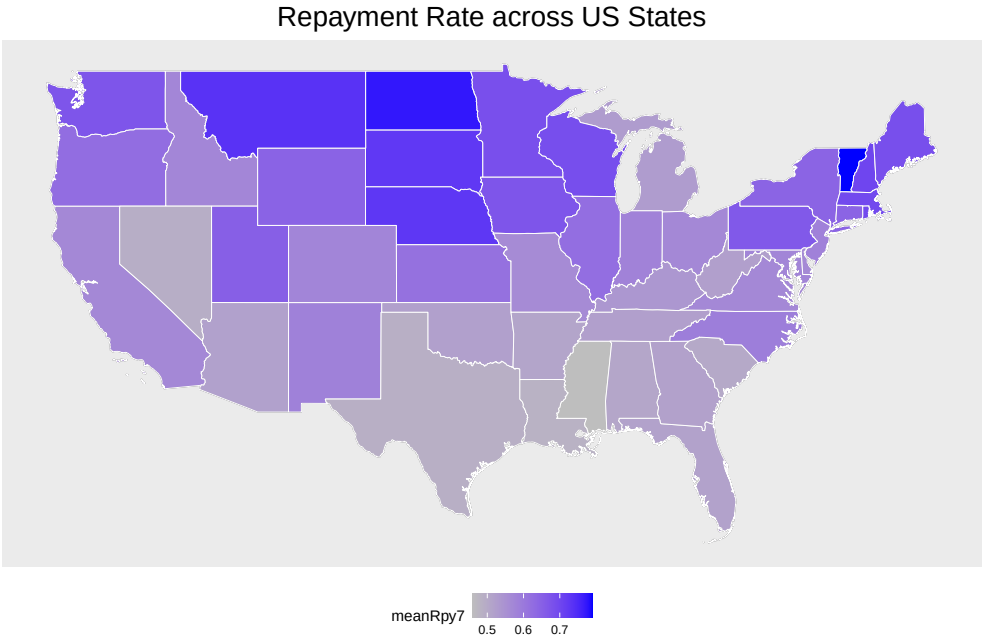
```
for (i in c(17,20:23)){
  print(names(studentDebtFinal)[i])
  print(table(studentDebtFinal[,i], useNA = "always"))
}
```

```
## [1] "STABBR"
##
##   AK   AL   AR   AZ   CA   CO   CT   DC   DE   FL   GA   GU   HI   IA   ID   IL
##    7   64   58   89  442   81   67   17   15  243  125    1   20   76   28  187
##   IN   KS   KY   LA   MA   MD   ME   MI   MN   MO   MS   MT   NC   ND   NE   NH
##  121   79   75   82  149   72   34  140  119  134   42   19  129   20   43   34
##   NJ   NM   NV   NY   OH   OK   OR   PA   PR   RI   SC   SD   TN   TX   UT   VA
##  117   40   31  305  257   77   73  329   56   20   84   22  114  325   44  121
##   VI   VT   WA   WI   WV   WY <NA>
##    2   22   85   80   49   11    0
## [1] "HCM2"
##
##    0    1 <NA>
## 5040   36    0
## [1] "PREDDEG"
##
##    0    1    2    3    4 <NA>
##  305 1752 1131 1862   26    0
## [1] "CONTROL"
##
##    1    2    3 <NA>
## 1597 1368 2111    0
## [1] "CURROPER"
##
##    1 <NA>
## 5076    0
```

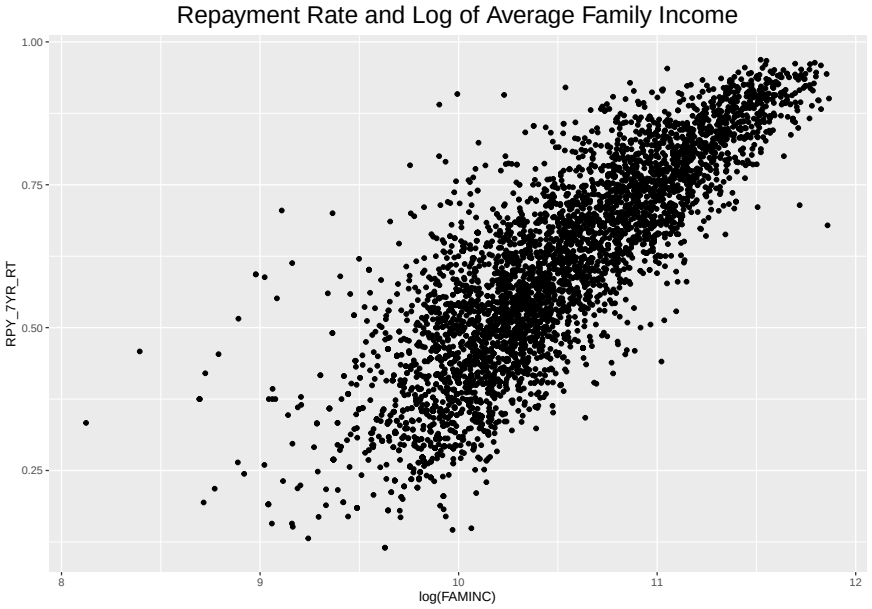
From the frequency tables, it is important to mention that CITY is not displayed because it has 2,003 different categories. In the case of states, they seem to be somehow representative of the US population density, for the states with the highest number of higher education institutions are California, Pennsylvania, Texas, and New York. Also, taking into account the dictionary of variables, PREDDEG shows that the majority of institutions offer predominantly Bachelor's Degrees, closely followed by Certificate Degrees, Associate's Degrees, Not Classified Degrees, and Graduate Degrees. Finally, the frequencies from the CONTROL variable are very interesting because 2,111 institutions are private for-profit, 1,597 are public and 1,368 are private nonprofit.

Visualization

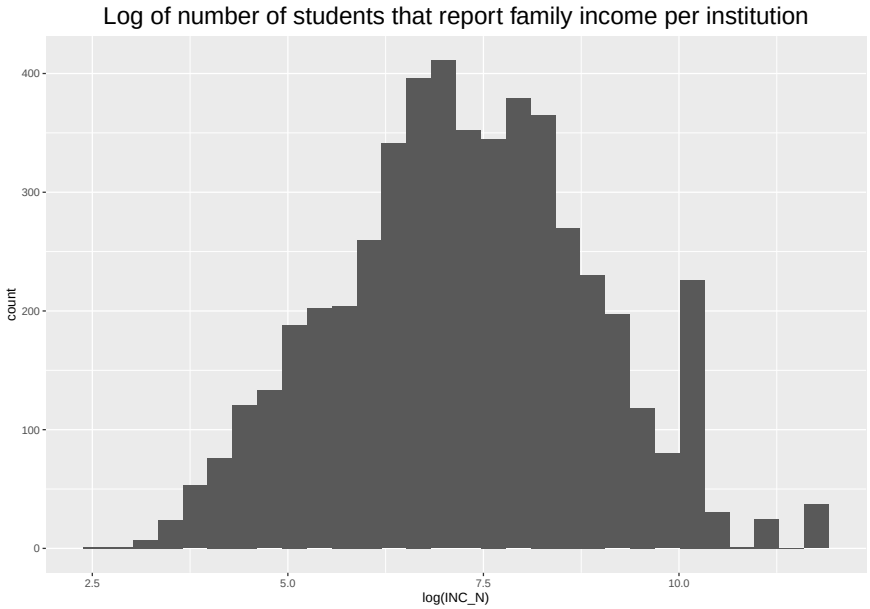
In terms of visualization, the three most important variables highlighted in this section are STATE, FAMINC and INC_N. The reason behind the selection of these three variables is based on the fact that they are the most representative variables to reflect demographic information of higher education students in the filtered dataset. For instance, Candidly which is the leading AI financial management tool for student-debt management in the US, mentions that 75% of its clients are women and people of color (Candidly, 2023). Thus, geographic location and family income may capture households of racial minorities who may have a lower repayment rate, or be largely concentrated in certain states with a lower average income. The first graph is a heat map which portrays the relationship between repayment rate and state across the US.



The map clearly shows that most northern states have higher average repayment rates compared to the states in the South. The following graph shows the relationship between repayment rate and the log of average family income.



From the results of the previous graph, there is a clear linear relationship between the repayment rate and the log of family income. The final graph is intended to display a histogram showing the distribution of the number of students that report family income for each institution.



The last graph shows that the distribution of the log of the number of students that report family income per institution follows a normal distribution.

Insights

From the visualization analysis, it is important to recognize that STATE, FAMINC and INC_N, seem to have a strong relationship with repayment rate, so those variables can be used to predict repayment rate. However, other variables such as AGE_ENTRY, PREDDEG, CONTROL, which have been identified from the filtering dataset can also add some predictive power to the model.

Modeling Approach

To further explore machine learning algorithms from the caret package in R, the "glmboost" algorithm was selected since it can process regression analysis by optimizing arbitrary loss functions where component-wise linear models are used as base learners (R Documentation, 2023). The caveats from the algorithm construction require that all numeric variables be scaled or centered, and that all categorical variables be transformed into factor variables. Based on the insights section, two models are identified. The first model will include STATE, FAMINC and INC_N as predictor variables, while the second model will include STATE, FAMINC, INC_N, AGE_ENTRY, PREDDEG, and CONTROL as predictor variables. For both models, the glmboost algorithm will be used to predict the repayment rate for the cohort of students who entered repayment 7 years ago.

In order to run both models, a simple data partition is employed with a 20% probability to select the test dataset. The train dataset which encompasses 80% of the data is used to estimate the model parameters, which will then be used to predict the rate of repayment on the test dataset. In both models, no cross-validation, further model resampling, or tuning parameters are used since the original dataset does not contain a very large number of observations, and the glmboost algorithm already runs internal optimization processes. It is important to note that the root-mean-square-error (RMSE) is used as a performance metric to compare how well both models performed on the test dataset. Finally, it is important to mention that FAMINC and INC_N incur a log-transformation in both models to capture the behavior displayed in the visualization section.

Section 2: Results

Upon running both models, it is necessary to mention that states with fewer than 30 institutions registered in the final dataset were deleted, because they would create noise on the performance metrics by using insufficient datapoints to estimate repayment. Although 41 states remained from the deletion, the number of observations in the dataset decreased minimally from 5.076 to 4,872. Also, the glmboost algorithm requires that the same states are included in both the train and test dataset, so that it can calculate the corresponding estimates and subsequently use them to predict repayment. Thus, eliminating those underrepresented states allowed the algorithm to run correctly and avoid additional bias on the estimates of each model.

In relation to the results obtained, the first model performed similarly well to the second model for they have a very small difference between their corresponding RMSE performance metrics. As a result, the second model which includes RPY_7YR_RT as a predicted variable and STATE, FAMINC and INC_N as predictors is selected, since its performance is more efficient by using fewer variables than the second model. In fact, the RMSE obtained for the first model has a value of 0.0870604 while the performance metric for the second is 0.0866926, which implies that the model predicts repayment rate with an average error of 8.7 percentage points.

The following table summarizes the performance metrics of both models employed:

method	RMSE
Model 1: GLMBOOST - repayment on income and state	0.0870604
Model 2: GLMBOOST - repayment on income, state, age, school financing, and main degree type	0.0866926

Section 3: Conclusion

This project allowed us to have a very generic overview of the big problem regarding the accruing student debt in the United States. Even though the datasets used came from cleaned data from Kaggle, it is useful to mention that the Department of Education periodically promotes very comprehensive data through the NSLDS and Scorecards, in order to motivate prominent researchers to explore the factors that affect student-debt repayment. The data cleaning approach used in this project was aimed at reducing the number of variables as much as possible, by particularly evaluating the completeness of the data based an a 0.5% acceptable treshold of missing values. Nonetheless, many variables which may have strong predictive power were excluded as a result, which should certainly be considered in future research.

Even though the data is aggregated at the institutional-level, the model results still portray the rethoric between income inequality in the United States and the ability for individuals to handle their financial obligations. The financial pressure that many people have to consider when going to college, is extremely controversial taking into account that many countries around the world offer free higher education. Further research should also focus on analyzing trends of student-debt over time, which can still be done through the data from the Department of Education which holds records starting from 1996. The results from future research can contribute to better approach this problem, which currently prevents many US citizens from accessing other financial opportunities such as buying a house or saving for retirement.

Section 4: References

Department of Education. (2016, March 2). College Performance, Debt and Earnings. Kaggle. https://www.kaggle.com/datasets/thedevastator/unlock-college-performance-debt-and-earnings-out/data

Department of Education. (2023, Dic 2). Data Documentation. Department of Education. https://collegescorecard.ed.gov/data/documentation/

Candidly (2023, Dic 2). Candidly works for your workplace. Candidly. https://getcandidly.com/employers/

R Documentation (2023, Dic 2). glmboost: Gradient Boosting with Component-wise Linear Models. DataCamp. https://www.rdocumentation.org/packages/mboost/versions/2.9-8/topics/glmboost

The Economist. (2020, Feb 22). Student debt in America amounts to over $1.5 trn. The Economist. https://www.economist.com/finance-and-economics/2020/02/22/student-debt-in-america-amounts-to-over-15trn?utm_medium=cpc.adword.pd&utm_source=google&ppccampaignID=18798097116&ppcadID=&utm_campaign=a.22brand_pmax&utm_content=conversion.direct-response.anonymous&gclid=CjwKCAiAxreqBhAxEiwAfGfndEgZshG3yyV4U6q55i8wpWUUZcMezm-QrHZDMxlWTgsXbzMAP_lR4hoCzUsQAvD_BwE&gclsrc=aw.ds

The World Bank. (2023, Dic 2). Data Bank. World Development Indicators. The World Bank: https://databank.worldbank.org/reports.aspx?source=2&series=NY.GDP.MKTP.CD&country=WLD

Welding Lyss (2023, Sep 1). How Long Does It Take To Pay Off Student Loans?. Best Colleges. https://www.bestcolleges.com/research/how-long-to-pay-off-student-loans/#:~:text=Data%20Summary,a%2010%2Dyear%20repayn