A predictive model to better evaluate the student loan default risk

Sanatan Das

April 08, 2018

Contents

The Problem: Student Loan Default Crisis

Summary of the Problem

This report analyzes new data on student debt and repayment, released by the U.S. Department of Education in October 2017. Previously available data have been limited to borrowers only, follow students for a relatively short period (3-5 years) after entering repayment, and had only limited information on student characteristics and experiences. The new data allow for the most comprehensive assessment to date of student debt and default from the moment students first enter college, to when they are repaying loans up to 20 years later, for two cohorts of first-time entrants (in 1995-96 and 2003-04). This report provides a broader perspective on student debt and default that considers all college entrants rather than just borrowers, provides substantially longer follow-up, and enables a more detailed analysis of trends over time and heterogeneity across subgroups than previously possible.

Key findings from new analysis of these data include:

- Trends for the 1996 entry cohort show that cumulative default rates continue to rise between 12 and 20 years after initial entry. Applying these trends to the 2004 entry cohort suggests that nearly 40 percent may default on their student loans by 2023.
- The new data show the importance of examining outcomes for all entrants, not just borrowers, since borrowing rates differ substantially across groups and over time. For example, for-profit borrowers default at twice the rate of public two-year borrowers (52 versus 26 percent after 12 years), but because for-profit students are more likely to borrow, the rate of default among all for-profit entrants is nearly four times that of public two-year entrants (47 percent versus 13 percent).
- The new data underscore that default rates depend more on student and institutional factors than on average levels of debt. For example, only 4 percent of white graduates who never attended a for-profit defaulted within 12 years of entry, compared to 67 percent of black dropouts who ever attended a for-profit. And while average debt per student has risen over time, defaults are highest among those who borrow relatively small amounts.
- Debt and default among black college students is at crisis levels, and even a bachelor's degree is no guarantee of security: black BA graduates default at five times the rate of white BA graduates (21 versus 4 percent), and are more likely to default than white dropouts. Trends over time are most alarming among for-profit colleges; out of 100 students who ever attended a for-profit, 23 defaulted within 12 years of starting college in the 1996 cohort compared to 43 in the 2004 cohort (compared to an increase from just 8 to 11 students among entrants who never attended a for-profit).
- Trends over time are most alarming among for-profit colleges; out of 100 students who ever attended a for-profit, 23 defaulted within 12 years of starting college in the 1996 cohort compared to 43 in the 2004 cohort (compared to an increase from just 8 to 11 students among entrants who never attended a for-profit).

The results suggest that diffuse concern with rising levels of average debt is misplaced. Rather, the results provide support for robust efforts to regulate the for-profit sector, to improve degree attainment and promote income-contingent loan repayment options for all students, and to more fully address the particular challenges faced by college students of color.

Background and data

Until recently, the dominant focus of public concern around student loans has been simply how much of it there is, and how rapidly it has been growing over time. At nearly $1.4 trillion in loans outstanding, student debt is now the second-largest source of household debt (after housing) and is the only form of consumer debt that continued to grow in the wake of the Great Recession.

But as many observers have noted, these aggregate statistics tell us little about the student-level experience with college debt. About one-quarter of the aggregate increase in student loans since 1989 is due to more students enrolling in college. More recent work that tracks debt outcomes for individual borrowers documents that the main problem is not high levels of debt per student (in fact, defaults are lower among those who borrow more, since this typically indicates higher levels of college attainment), but rather the low earnings of dropout and for-profit students, who have high rates of default even on relatively small debts.

This study utilizes new data, released by the U.S. Department of Education in October 2017, linking two waves of the Beginning Postsecondary Student (BPS) survey, a nationally-representative survey of first-time college beginners, to administrative data on debt and defaults. This allows for the most comprehensive assessment yet of student debt and default from the moment students first enter college, to when they are repaying loans up to 20 years later, for two cohorts of first-time entrants (1995-96 and 2003-04 entrants, which I refer to as the BPS-96 and BPS-04 as shorthand).

This allows for a broader perspective that considers all first-time college entrants rather than just borrowers, provides substantially longer follow-up than other data sources, and enables a more detailed analysis of trends over time and heterogeneity across subgroups.

How default evolve over time, by entry Cohort

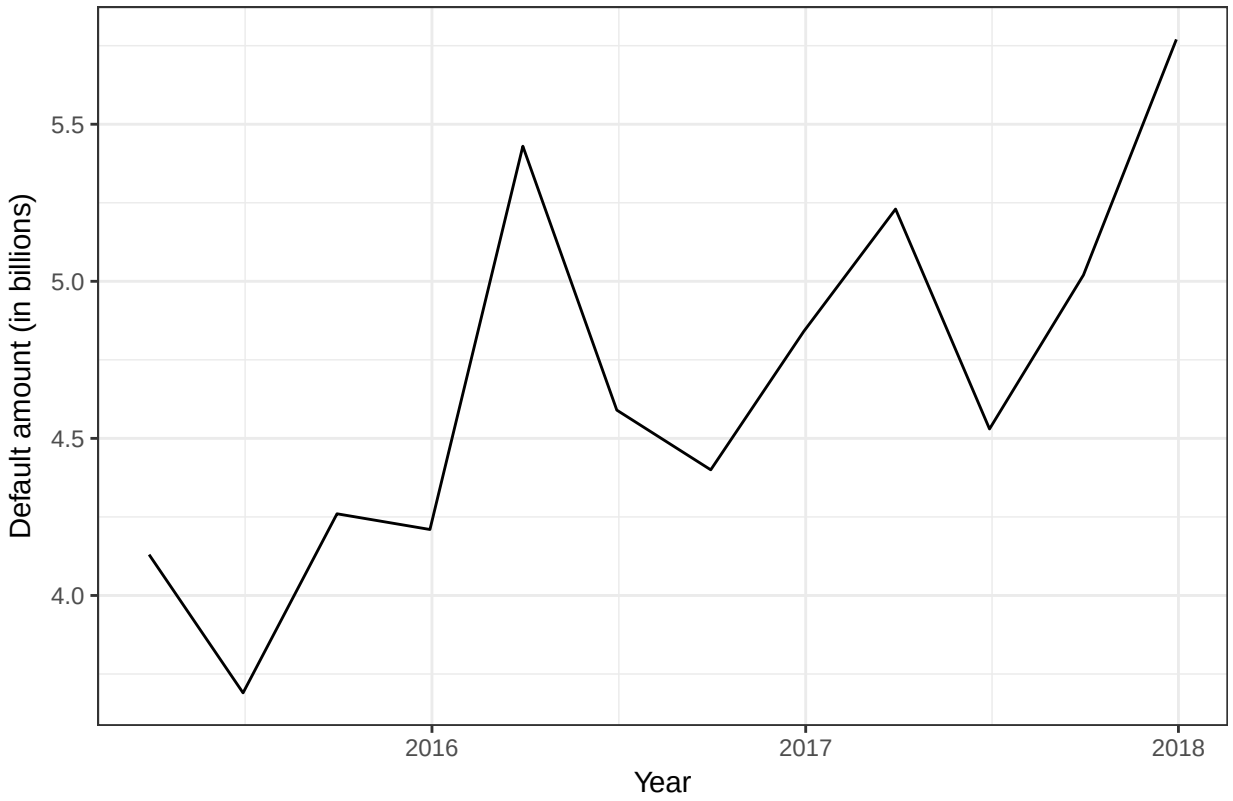
The best prior estimates of overall default rates come from Looney and Yannelis (2015), who examine defaults up to five years after entering repayment, and Miller (2017), who uses the new BPS-04 data to examine default rates within 12 years of college entry. These two sources provide similar estimates: about 28 to 29 percent of all borrowers ultimately default.

But even 12 years may not be long enough to get a complete picture of defaults. The new data also allow loan outcomes to be tracked for a full 20 years after initial college entry, though only for the 1996 entry cohort. Still, examining patterns of default over a longer period for the 1996 cohort can help us estimate what to expect in the coming years for the more recent cohort.

If we assume that the cumulative defaults grow at the same rate (in percentage terms) for the 2004 cohort as for the earlier cohort, we can project how defaults are likely to increase beyond year 12 for the 2004 cohort. To compute these projections, I first use the 1996 cohort to calculate the cumulative default rates in years 13-20 as a percentage of year 12 cumulative default rates. I then take this percentage for years 13-20 and apply it to the 12-year rate observed for the 2004 cohort. So, for example, since the 20-year rate was 41 percent higher than the 12-year rate for the 1996 cohort, I project the Year 20 cumulative default rate for the 2004 cohort is projected to be 41 percent higher than its 12-year rate.

Figure 1 plots the resulting cumulative rates of default of the last three years data (Data Source : Federal Student Aid)

Figure1 : Cumulative Lifetime Default Rates



Based on the patterns observed for the earlier cohort, a simple projection indicates that about 38 percent of all borrowers from the 2003-04 cohort will have experienced a default by 2023.

Conclusion

The analyses presented above highlight the value of tracking individual students from the beginning of their college trajectory for many years beyond when they leave school, and the importance of disaggregating trends by student and institutional characteristics. Key findings include:

- Trends for the 1996 entry cohort show that cumulative default rates continue to rise between 12 and 20 years after initial entry. Applying these trends to the 2004 entry cohort suggests that nearly 40 percent may default on their student loans by 2023.
- The new data show the importance of examining outcomes for all entrants, not just borrowers, since borrowing rates differ substantially across groups and over time. For example, for-profit borrowers default at twice the rate of public two-year borrowers (52 versus 26 percent after 12 years), the rate of default among all for-profit entrants is nearly four times that of public two-year entrants (47 percent versus 13 percent).
- The new data underscore that default rates depend more on student and institutional factors than on average levels of debt. For example, only 4 percent of white graduates who never attended a for-profit defaulted within 12 years of entry, compared to 67 percent of black dropouts who ever attended a for-profit. And while average debt per student has risen over time, defaults are highest among those who borrow relatively small amounts.
- Debt and default among black or African-American college students is at crisis levels, and even a bachelor's degree is no guarantee of security: black BA graduates default at five times the rate of white BA graduates (21 versus 4 percent), and are more likely to default than white dropouts.

- Trends over time are most alarming among for-profit colleges; out of 100 students who started college at a for-profit, 23 defaulted within 12 years of starting college in the 1996 cohort compared to 43 in the 2004 cohort (compared to an increase from just 8 to 11 students among entrants who never attended a for-profit).

To conclude, the results suggest that diffuse concern with rising levels of average debt is in different areas. A number of factors are involved in the student loan to be default.

Now, it is the responsibility of the Banks and Credit Unions to do a thorough analysis on the loan application and the student data before approving the loan.

References

Statistical Reports -

- https://www.brookings.edu/
- https://www.insidehighered.com/quicktakes/2018/01/12/new-analysis-student-loan-default-data
- https://www.forbes.com/sites/zackfriedman/2017/10/06/student-loan-default/#405c62f028de

Database -

- https://studentaid.ed.gov/sa/about/data-center/student/default
- https://www2.ed.gov/offices/OSFAP/defaultmanagement/instructions.html

The Solution: Educational Loan Assistant

What is 'Educational Loan Assistant'?

The Educational Loan Assistant is a tool that provides guidance and analysis to the **Banks and Credit Unions** before they approve any educational loan to the student loan applications.This tool provides a thorough analysis report on the information provided in the loan application based on the predictive model we build. It also depicts the likelihood of a loan application being default (Not possible, May be possible, Strongly possible, Not recommended). It is the final decision of the Bank and the Credit Union agents to take on the loan approval. The Educational Loan Assistant is a helper to report the risks so that the banker or lender can take well informed decision.

Every year U.S. Department of Education releases a detailed dataset on the student debt and repayment. These datasets have wide variety of features and information about the educational loan system in the USA. This project does a thorough analysis of those data and builds a robust predictive model for predicting the risk on student loan approvals to help the lender organizations. This porjet also buils an API (application programming interface) and a web based tool, so that the bankers and lending agents can use it. The API is built in such a way that any developer can use it from different kind of client applications (web apps or mobile apps)

This tool will be able to train and build the model again and again when new data set is added to the existing data set as every year a new data set will be released by the U.S. Department of Education.

It will also generate a monthly report on the predictions and a yearly report of validations to validate if there is any default in a student loan where the tool predicted as no risk or low risk. Althouh we have to wait for some time to see the first analysis/validation report (tool performace metrics).

The Project Requirements

Goal of this project

The goal of this project is to create a predictive model that will help the Lending Organizations (student loan) to better evaluate the student loan applications risks of being default. The target of this project is to achieve >95% correct prediction of the default rate. Although this will take some time to evaluate the result.

Business Understanding (Problem Formulation)

The Banks and Credit Unions spends billions of dollars every year for recovering the student loan debt. Feferal Student Aid releases multiple daasets and reports every for the Official Cohort Default Rates for Schools. These datsets are available in their website for download in different formats (excel, csv, html and pdf files). They also publish the detailed instructions how to use these data files. They also publish the field definitions, abbreviations used etc. so that the values can be extracted from those information. To solve this problem of the lending organizations, we analyze the data and create a model to help them to reduce the defaults.

Data Collection

The first step for this project is to collect those individual files and go through them, looking for the important information. Then we need to put them together into a single format (csv). The data cleaning, verifying the value types (char, numeric etc) need to be completed by looking at the data (before we start the thorough cleaning and making it tidy using R).

Exploratory Data Analysis, Visualization

Once the dataset is ready, we need to start on the Exploratory Data Analysis, creating visual representations, plotting relationship graphs, finding strong and weak features, finding outliers etc. The whole exercise willproduce a set of features (a broader set) that will be used for creating models.

The dataset analysis and feature extractions examples:

Default count by School Type (Year 2014)

SchoolType	DefaultCount
Public	1571
Private	1477
Nonprofit	1503
Proprietary	125
Foreign public	31
Foreign private	5
Foreign For-Profit	1

Default count by Program Length (Year 2014)

ProgLength	DefaultCount
Short-Term (300–599 hours)	1
Graduate/Professional	1

ProgLength	DefaultCount
Non-Degree (600–899 hours)	69
Non-Degree 1 Year (900–1799 hours)	872
Non-Degree 2 Years (1800–2699 hours)	239
Associates Degree	1085
Bachelors Degree	504
First Professional Degree	35
Masters Degree or Doctors Degree	1879
Professional Certification	0
Undergraduate (Previous Degree Required)	0
Non-Degree 3 Plus Years	26
Two-Year Transfer	1

Default count by Race/Ethnicity (Year 2014)

Ethnicity	DefaultCount
Native American	4
HBCU	94
Hispanic	143
Traditionally Black College	4471
Ethnicity Not Reported	1

Model Training/Analytics

When the feature set is ready we will start creating the models and start analyzing the statistical parameters like coefficients, standard error, R-squared, p-values, ROC and AUC values. That will help us finalizing the feature set and the final model we choose. Once the model is ready, we will use the validation set and test set to verify our results and verify with the business.

Deployment and the Tool

We will build a Java based web tool to run the model. The User interface will be able to provide the input data for student loan application and the tool (API) will run the model in the background and predict the risk and likelihood of being default of that loan application.

User Acceptance

Finally we will run the aplication for agents to help them and generate some sample reports (monthly and annually)

What are the benefits from the solution (predictive modeling)?

There will be a list of advantages that the Bank and Credit Unions will get while using this tool (model):

- **A recommendation engine** : The phone banker, store banker and the lending agents will get a full insght of information, risk and statistcal advice from this tool before approving any student loan. This will reduce the number of default incidents in a significant number for the lending organizations.

- **Reduce huge monetary loss** : The above statistical report shows that increasing number of student loan default is a threat and huge loss for Banks, Credit Unions or any Lending organizations. We see that billions of dollars are not paid or not paid on time in the student loan business. Any prevention of the default rate will save a significant amount of money of the leding organizations. ***The success criteria of this model is to predict the default risk by > 95% true prediction.***
- **Business growth** : The tool/model will generate a monthly and annual report so that they can measure the business performance metrics.
- **New product ideas** : The product managers and the business analysts will get to know the important factors of the Educational Loan business. They might get inspired to revise their list of information that they collect from the student/family/individuals through paper based applications or electronically.
- **Improving global economy** : The decreasing number od defaults in the lending business will strengthen the national and global economy and reduce the chances of recession. Overall economicall growth will increase.

Who are going to use this solution (business users)?

This tool (model) can be used by multiple business users (of different roles) for their own benefits in the Banking and Student Loan organizations. We can think of the follwing business users who can use this model (in a different format: Web Application or PDF Reports) and get different kinds of benefits out of it.

- **A phone banker/store banker or lending agent** : This kind of users will use the user interfaace via a website or a web application to get the instant help on analysing the stundent loan application data and seek for a recommendation whether or not to approve a loan application. These are daily business users. Most of the users will fall into this category.
- **Bank, Credit Unions and Lending Organizations Management** : This kind of users will use the tool in the format of reports and metrics, will look at the figures and predict the future profit of the organization. They are not daily users. They will see the reports monthly an annually or on-demand basis.
- **Business Analysts and Product Managers** : This kind of users will get an experience of both the formats (Web Application and the PDF Reports). They will use those to create new business ideas, improving their existing products and creating new products.

Timelines of the solution development

The high level estimated timelines for this project implementation will be as below:

Tasks	Timelines
Creating formal business requirements	1 week
Data Collection and Data Understanding	2 weeks
Exploratory Data Analysis	1 week
Model Training and Analytics	2 weeks
Deployment	1 week
User Acceptance Testing	1 week

Risks Involved (estimated)

The goal of this solution is to help the Student Loan organizations to make well informed decisions while approving any educational loan application. Currently today they are already having a high rate of the

default applications as mentioned in the report. It can't get worse than as it is today. But still this project and the model comes with some low severity risks with it. We can list down the risk statements clearly as below so that the business is aware of before delving into the model development.

- The proposed model is built on the data released by the U.S. Department of Education. It is assumed that those data are correct and provides the right trend of the student loan debt and default. The machine learning model is all based on the data we trust. If there is any issue on the provided data, it will impact the model and prediction behavior.
- As we have seen in the statistical reports that the student loan default parameters/features have changed time to time. So the model parameters need to be updated accordingly time to time, feature list needs to be updated too to make more accurate preditions.
- To get the first real time prediction validation (is the prediction adding value to business?), the business/management have to wait for a quite some time as the loan payment starts long after the approval and disbursement of the loan.

Although all these risks can be minimzed with proper training, validation and the test set of the data.

Solution deployment, post production maintenance and reports

The proposed model will be deployed under a wrapper of an web API so that the above mentioned user can use the model through a nice set of user interfaces. The batch job will run periidocally for generating PDF reports. The deployment model will look like below.

- **Application Module** - Model->JRI API (an API for Java and R integration)->Java Application->Web APPlication
- **Batch Module** - Model->JRI API (an API for Java and R integration)->Java Application->PDF Reports (email the reports or store in a directory location)

The post porduction and the maintenance team will be using the log statements and checking the model/tool performance metrics on regular basis.

Conclusion

The tool 'Educational Loan Assistant' analyzes the data based on student debt and repayment information, released by the U.S. Department of Education and creates a predictive model. To achieve more accuracy on the model prediction, we can use the private and confidential data of the Bank and Credit Unions (if we create a model specific to the organization and every lending organization ahs their customer database). Those are more realistic and accurate information to use and will provide better results. This idea can also be extended to other Lines of Businesses (LOBs) like **Home Mortgage Loans**, **Personal Loans and Lines of Credit**, **Auto Loans** etc. Although, every specific Lines of Business will have a different set of features/predictors to work with. If this model looks promising and the business users get significant benefit using this model, we will extend the same idea for other Lines of Businesses in future.