Project

Kelly Chen

02/12/2019

Contents

1 Introduction

With student loan debt skyrocketing in recent years, many families in the US are struggling to pay off the student loan. This project aims to provide information on both sides. For families that potentially want to take a student loan, they need a clear picture of how the student loan is going to affect the family financially in the future. What the financial struggles they are going through? For US government or policy-makers to make better policies to relieve the burden of student loans on family, they need to know what kind of families are more likely to have a high student loan amount, so they design a better plan to fit the families or adjust the bar for entering an affordable plan.

The project consists of two parts. The first part is finding which features of a family can predict the amount of student loan debt, among them which one has the biggest impact. The second part is to identify which financial hardship the education loan most likely leads to.

2 Problem Statement and Background

Affordability of student loans is a rising topic in the United States in recent years. The expense of education — tuition fee keeps rising every year but it does not stop Americans from borrowing money to go to school. Today, more than 44 million people carry over 1.5 trillion dollar of outstanding student loan debt, an amount that exceeds all other types of non-mortgage loan debt(A $1.5 Trillion Crisis: Protecting Student Borrowers and Holding Student Loan Servicers Accountable, 2019). While it is good to see that people attach more importance to education, and are willing to take a risk for academic advancement. However, many found themselves stuck in an unexpected financial hardship after getting a student loan. The Pew Charitable Trusts published a report on “Student Loan System Presents Repayment Challenges” in 2019, in which it analyzed 400,0000 people in Texas who are in student debt, and found out that many are in distress and are struggling to pay back the loan. Approximately a quarter of borrowers defaulted within five years of entering repayment(The Pew Charitable Trusts, 2019). The thesis “student loan debt and house household financial hardship: an analysis using the 2016 survey of consumer finances” concluded that student loan contributes to household financial hardship, using the same dataset this project uses(AlQaisi & Kern, 2018). This project, however, focuses on the correlation between family features and student loan amount, and the correlation between the student loan amount and financial hardship indicators.

3 Data

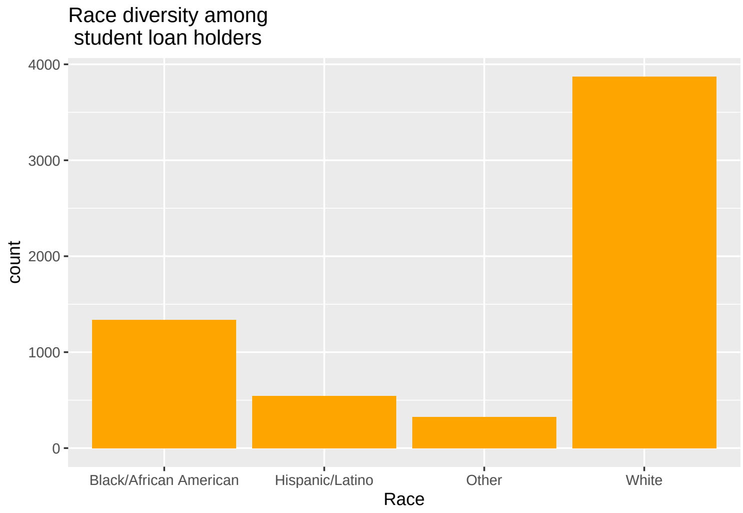
The data for this project comes from The Survey of Consumer Finances in 2016 (which can be downloaded here https://www.federalreserve.gov/econres/scfindex.htm). The Survey of Consumer Finances(SFC) is a dataset that contains all the data needed for the analysis and is provided by The Federal Reserve to help the government and ultimately the public to understand the financial condition of families in the United States. This data is based on a survey of six thousand households in the US about their finance. All the variables in the dataset are survey questions, the unit of observations is the US household. This dataset also comes with a codebook, which can also be downloaded on the same website. In total, there are 31240 observations and 5320 variables. For the first part of the project, I chose 15 variables as predictors of the student loan amount. They are: Highest received education, Confidence in the US economy, Knowledge about own finance, Loan amount, Willingness to take financial risk, How many loans, Owe money to purchase property, Trust fund, Main bank account balance, Income, Life insurance, Balance of saving account, Value of certificate of deposit, Value of saving bonds, Market value of stock neutral funds. These are the financial indicators of a family. Additionally to predictors, basic demographic characteristics variables are also selected, including Age, Race, Gender of the head of the family, Marital status and Kind of housing. These variables are not used as predictors but the relationships between them and student loan amounts are graphed. For the second part of the project, I chose 6 variables to present a financial hardship. They are Loan amount, Credit denied, Late debt payment, Payday loan, Bankruptcy, Foreclosure. These variables are chosen based on the belief that they can well represent the financial hardship that a family faces.

When it comes to data wrangling, most of the work focused on interpreting the codebook. According to the codebook, I selected the above variables, rename all the columns as their original meanings. For categorical values, I also renamed them based on their meanings. Some answers in the survey are not applicable or answered with “don’t know/not sure”, and they were given the value of 0 in the dataset. These values are treated as missing value for the analysis and are turned into N/A during the data wrangling stage.

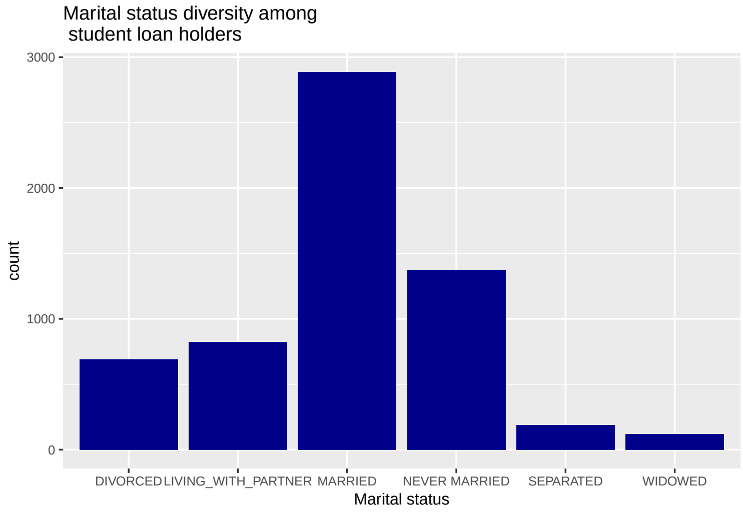
4 Analysis

The first step for analysis is to find the relationship between the dependent variable (student loan amount) and independent variables. For categorical variables such as race, marital status, Highest received education, which cannot be assigned to a certain level of loan amount relevance, I graphed their relationships to give a clear illustration of the difference of student loan amount between different groups.

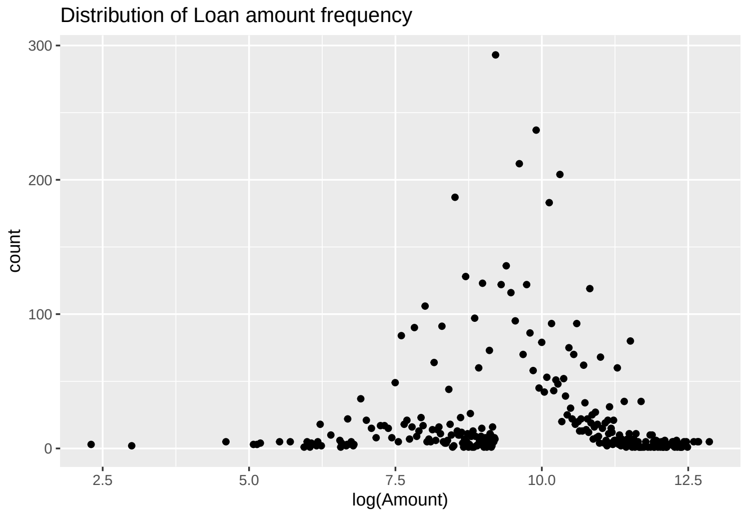
```
race_plot
```



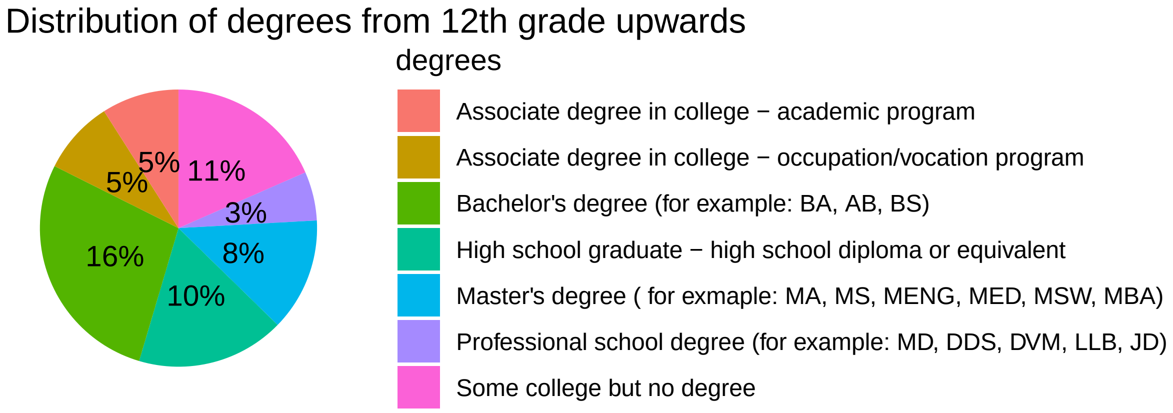
```
marital_plot
```



```
loan_frequency
```

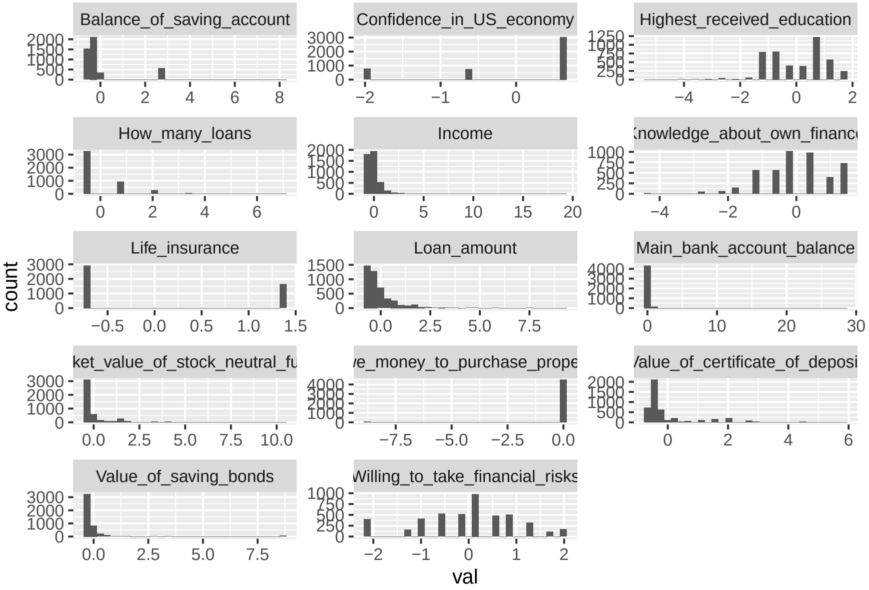


```
pie
```



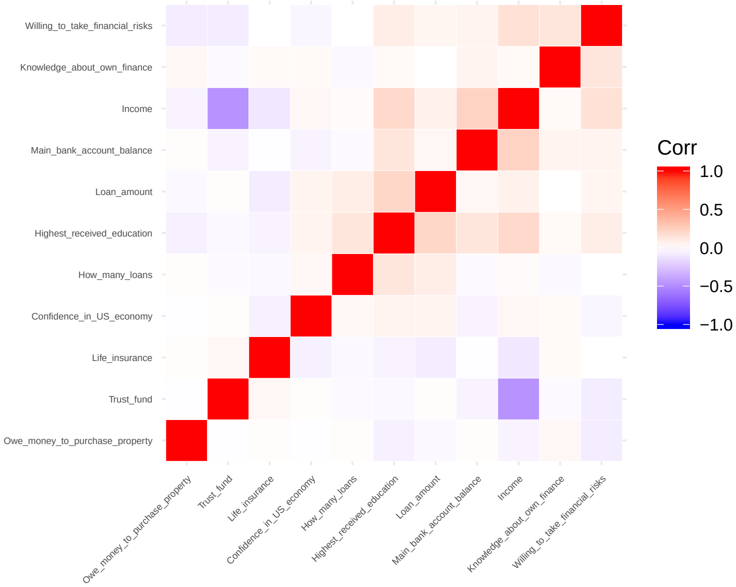
The second step is to prepare the dataset for machine learning, which means I partition the dataset into the training set and the testing set. The training set is 75% of the data, which I use to investigate and train models. The testing set has 25% of the data, which I use to test out the accuracy of the model prediction. In order to find the best fitting model, I first preprocess the training data. The purpose of data preprocessing is to scale and transform the continuous variables, so the data can fall within the same numerical range; to impute the missing data for better-estimated results; to turn categorical variables into dummy variables to fit into the model in the next step. With the help of Cran package, a framework for creating and preprocess design matrices, I transform the variables to have a mean of 0 and variance of 1; impute the missing data and transformed categorical data. These changes are also applied to both training data and testing data in order to calculate out of sample predictions. After preprocessing, we can see the distribution and variation of each variable as follows:

```
training2_plot
```



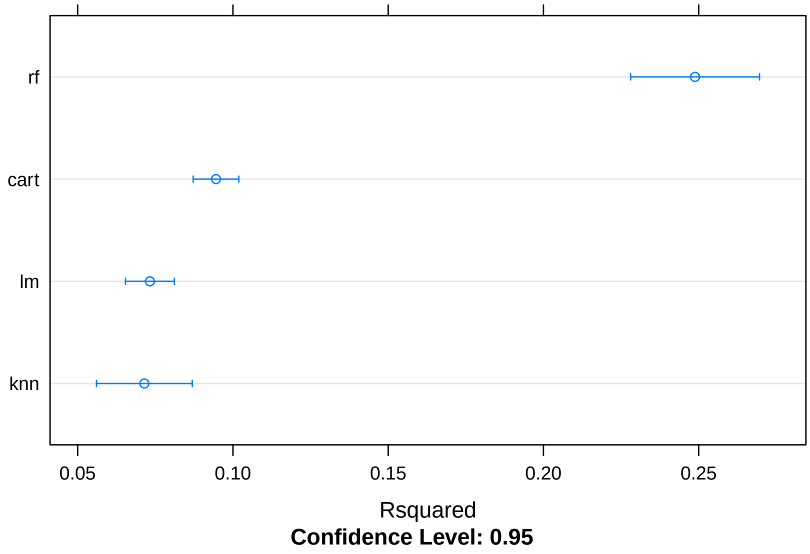
The third step is to establish the correlation between the dependent variable and independent variables. Here I plotted a correlation graph, from which we can see that the student loan amount and all the variables are correlated except for Trust Fund. Since there is no correlation between the Trust Fund and student loan amount, I dropped this variable.

```
corplot
```



Since we are predicting a continuous/interval-based outcome —— student loan amount, I decided to run 4 different regression models random forest, classification and regression trees, linear regression and K nearest neighbors, to test out, which model performs the best and can potentially be used to predict student loan amount using independent variables. After generating the models from training data, I calculated the R-square value in testing data using the models. From the result, we can tell that the Random Forest performs the best with the highest R-square value close to 0.25. R-square measures the goodness of fit of the model, the higher the R-square value is, the better the model fits.

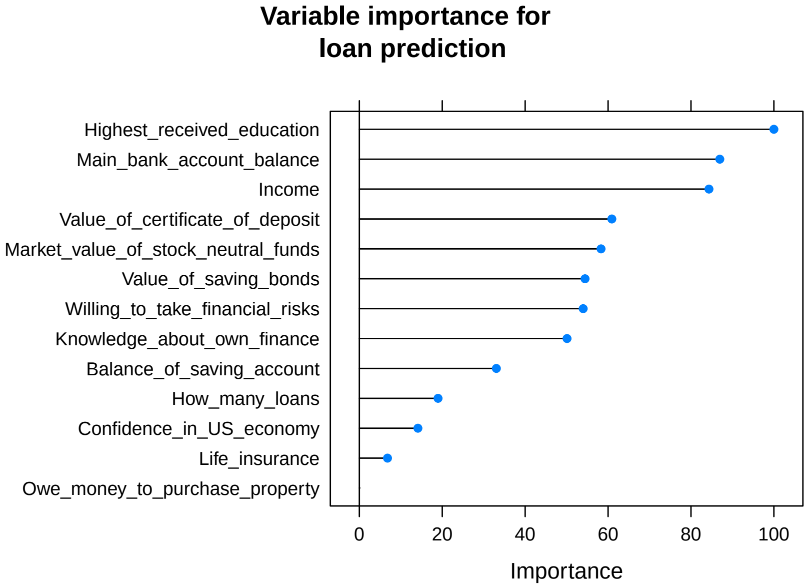
```
dotplot(resamples(mod_list1),metric = "Rsquared")
```



To validate that Random forest is the most-fitting model, I calculated the MSE of the model, MSE measures the average of the squares of the errors, the lower the MSE is, the less the errors are. The MSE of the Random forest is 0.1391496, which is close to 0, which means the model can well predict the student loan amount based on the variables.

To further investigate which variable weights more in terms of deciding the value of the dependent variable. I used the importance function from the random forest package to show how important each variable is to the student loan amount.

```
plot(importance_1, main="Variable importance for \n loan prediction")
```



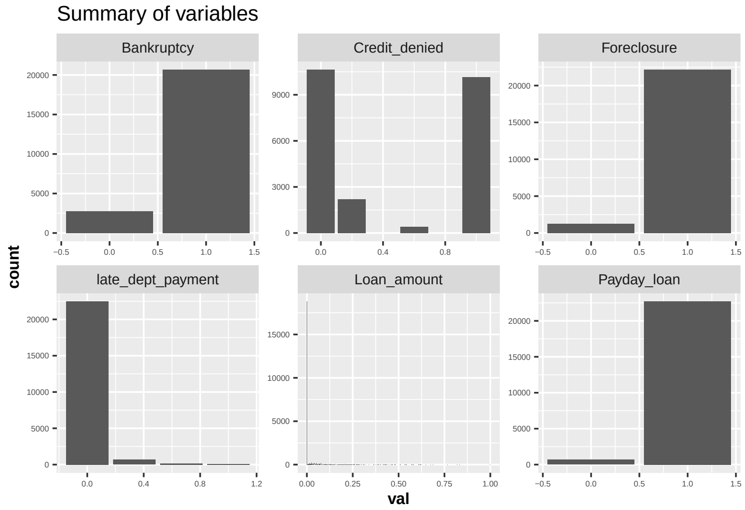
From the importance table, we can see that the variables Highest Received Education has the highest importance, followed by Main Bank Account Balance and Income which have similar importance.

For the second part of the project, I changed the independent variables to the Loan amount, Credit denied, Late debt payment, Payday loan, Bankruptcy, Foreclosure. All of the variables are categorical and consist of binary answers except for the credit denied variable and the loan amount itself.

Although here I use financial hardship indicators as independent variables here (X), and student loan amount as the dependent variable (Y), X is as correlated with Y as Y is with X. R-square is the same between regression of X on Y and of Y on X. Therefore, for any regression model that can represent the relationship between X and Y, it can also be considered how Y (student loan amount) is affecting X □financial hardship indicators).

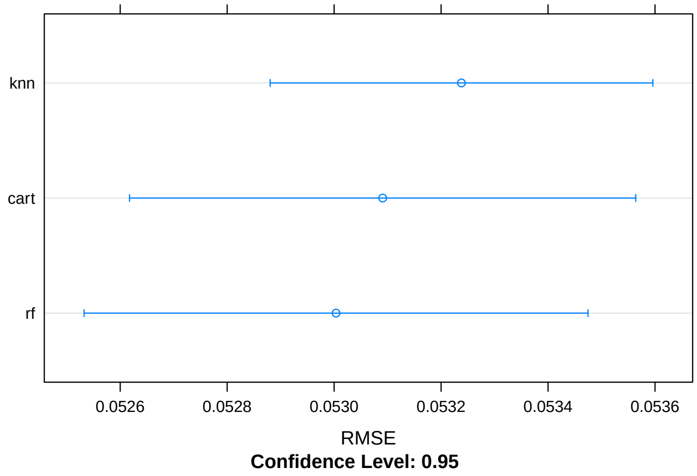
Follow the same stream in part 1, I used the same data partition and preprocessing process. The difference is, all the variables are categorical in part 2, so the distribution of each variable is also quite different. As it is shown in the following:

```
hardship_summary
```



In order to establish the relationship between student loan amount and financial hardship for future prediction, three models are also generated from training data and are tested in the testing data. The three models are Random Forest, Classification and Regression Trees. Linear regression is not chosen here for it is a parametric model, which has a strong assumption about the form f(x). Since all the variables are categorical here, linear regression will not be effective in predicting. After testing three models, RMSE is calculated and compared among them. s

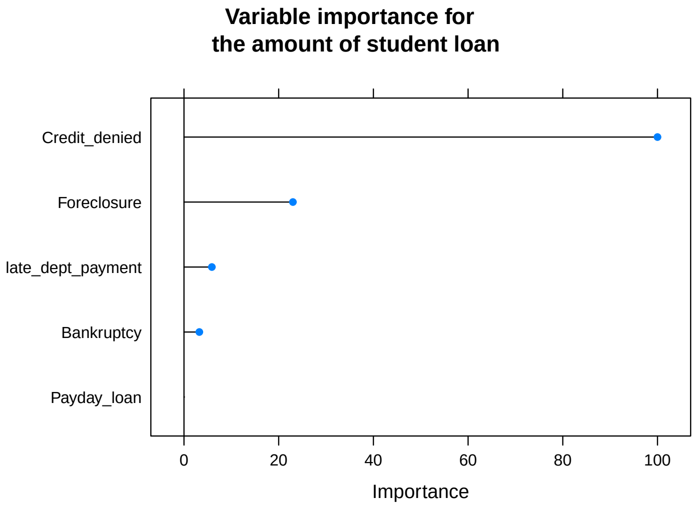
```
dotplot(resamples(mod_list2),metric = "RMSE")
```



From the result above, we can see that the random forest model again has the lowest RMSE, which means it has the least error when predicting. Therefore, the random forest model is the best-fitting model to predict the financial hardship type among them all. In addition, for a range between 0-1, a 0.053 RMSE score is really good for the model.

In the last part of the project, the importance function is used again to weigh the contribution of each variable makes to the student loan amount, or it can also be interpreted as the likelihood that these financial hardships appear based on the student loan amount.

```
plot(importance_2, main="Variable importance for \n the amount of student loan")
```



Here we can see that a higher student loan amount most likely leads to a credit denial, followed by foreclosure and late debt payment.

Result Summary In part 1, we can see that the distribution of student loan amount is skewed to the left, which means for people who get student loans, the amount on average is quite high. We also observed the highest proportion of people who have student loans is people with bachelor's degrees among different education levels (more than 50% of student loan holders have graduated from high school or higher educational institutions); married people among different marital statuses; while people among different races.

From the result of comparing R-squared value between four models, I concluded that random forest has the highest R-square value which means it is the best fitting model out of all four models. Also, the random forest model has a really low MSE of 0.1391496, which means we can choose this model to predict student loan amount using all the variables I chose. Out of all the variables, I also found that the highest received education is the most important variable, this means this variable has the highest influence on student loan compared to others. The higher the education people receive, the more likely the family is going to have a higher student loan amount.

In part 2, the random forest again performs the best when predicting financial struggles, with the lowest RMSE of 0.053, which is really good considering the scale of the data. From the importance table, I concluded that a higher student loan amount more likely results in credit denied than other situations. If a family has a high student loan debt, they would possibly end up facing credit denied, foreclose and late debt payment.

5 Reference

AlQaisi, R., & Kern, A. (2018). Student Loan Debt and Household Financial Hardship: Analysis Using the 2016 Survey of Consumer Finances (dissertation).

A $1.5 Trillion Crisis: Protecting Student Borrowers and Holding Student Loan Servicers Accountable: Hearings before the United States House Committee on Financial Services (2019) (Testimony of Ashely C. Harrington).

The Pew Charitable Trusts report. (2019). Student Loan System Presents Repayment Challenges. Retrieved from https://www.pewtrusts.org/-/media/assets/2019/11/psbs_report.pdf?la=en&hash=F9E369C81CB858FCE2A0B468087236CB3ACD65C6