Student Loan Repayment Prediction Report

July 19, 2017

Student Loan Repayment Analysis

Executive Summary

Higher education comprises a diverse range of colleges and universities that vary significantly terms of quality and cost, making it challenging to evaluate college performance and difficult for students and families to understand which college options are most suitable to them. As tuition costs rise, more and more students are taking out loans to cover the cost of their educations. This report, try to find out the repayment rate of student loan and to answer the question: just how good of an investment is higher education?

The analysis is based on 8705 student loan records published by the United States Department of education: https://collegescorecard.ed.gov/. The goal of this report is providing loans to attend what type of school makes it likely the student will graduate and get a job with enough income to repay those loans comfortably. This document has provided model and predict the percent of students repay their loans. After exploring the data, is quite evident that ability to repay loans depends on the student family circumstance, the school they went, the degree they earned, the job they get after graduation, and how much they borrowed?

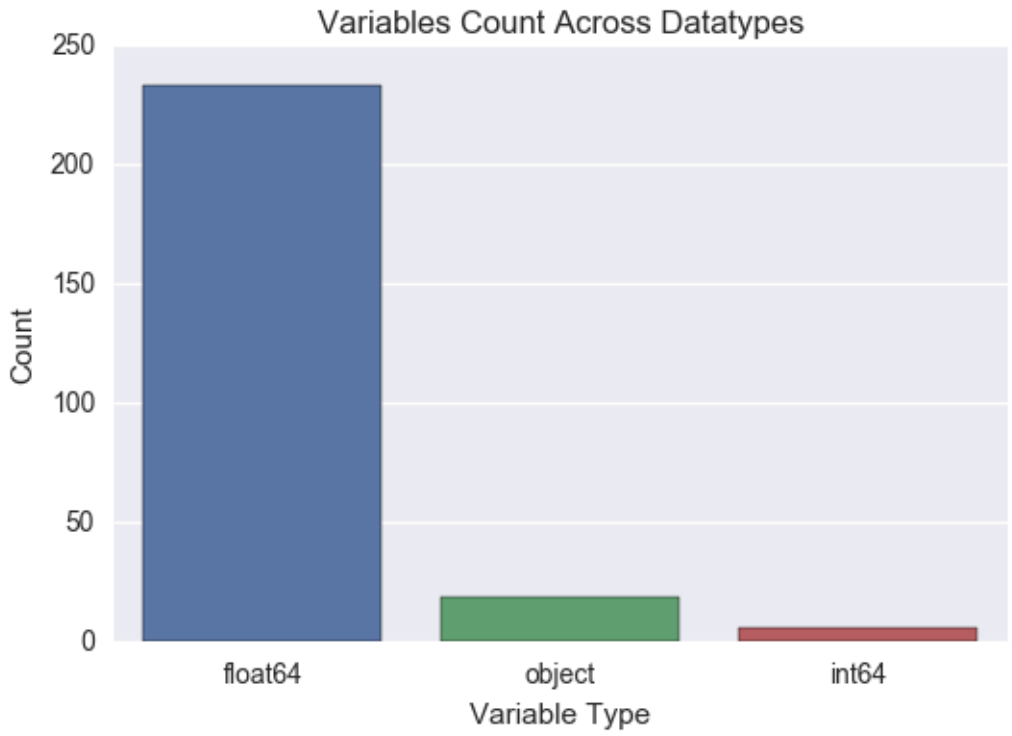
Although many factors can influence the ability to repay this report has found following are the key attributes:

- Student family income.
- Admission SAT scores.
- Quality of school
- Financial dependency status
- Parent Education level
- Cost of attendance, net price
- Level of degree offered
- Cumulative loan debt
- Access to federal aid (Pell grant)

Summary statistics and visualisations of the datasets, reveal potential relationships between loan repayment rates with students' family income, type of college they went, the ethnic background they come from, the admission sat score and income level they generate after graduation.

Data Exploration

The dataset has very large number of features and their association with repayment rate can be found out only by knowing more about their type. Let us visualize features counts in our dataset



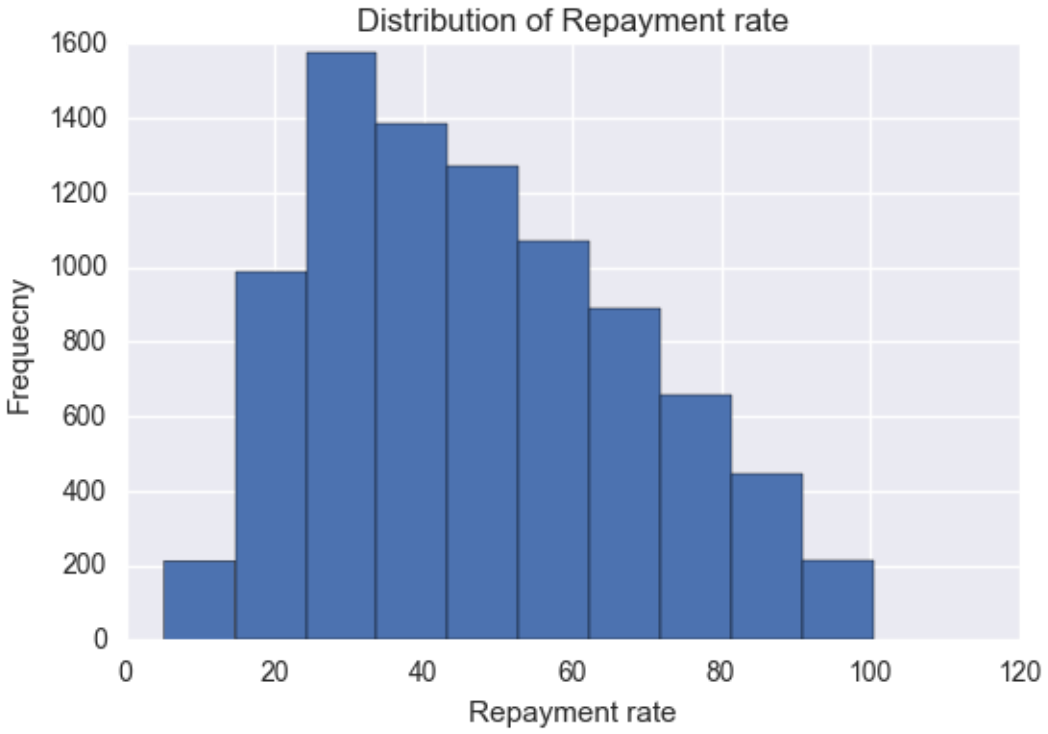
More than 400 numeric features in the data means there will be correlation of several features with repayment rates. Main objective of this data exploration to find the causation of this relationship. While most of these data do not necessarily reflect how a specific individual's outcome, this report offers exploratory analyses of how federal data may be used to measure an institution impact on a subset of performance measures. This report focuses on one outcome of higher education successful student loan repayment rate and do not attempt to define or measure the broader purposes of higher education.

Univariate analysis

Since Repayment rate is of the interest of this analysis, descriptive statistics shows the distribution of it in the data.

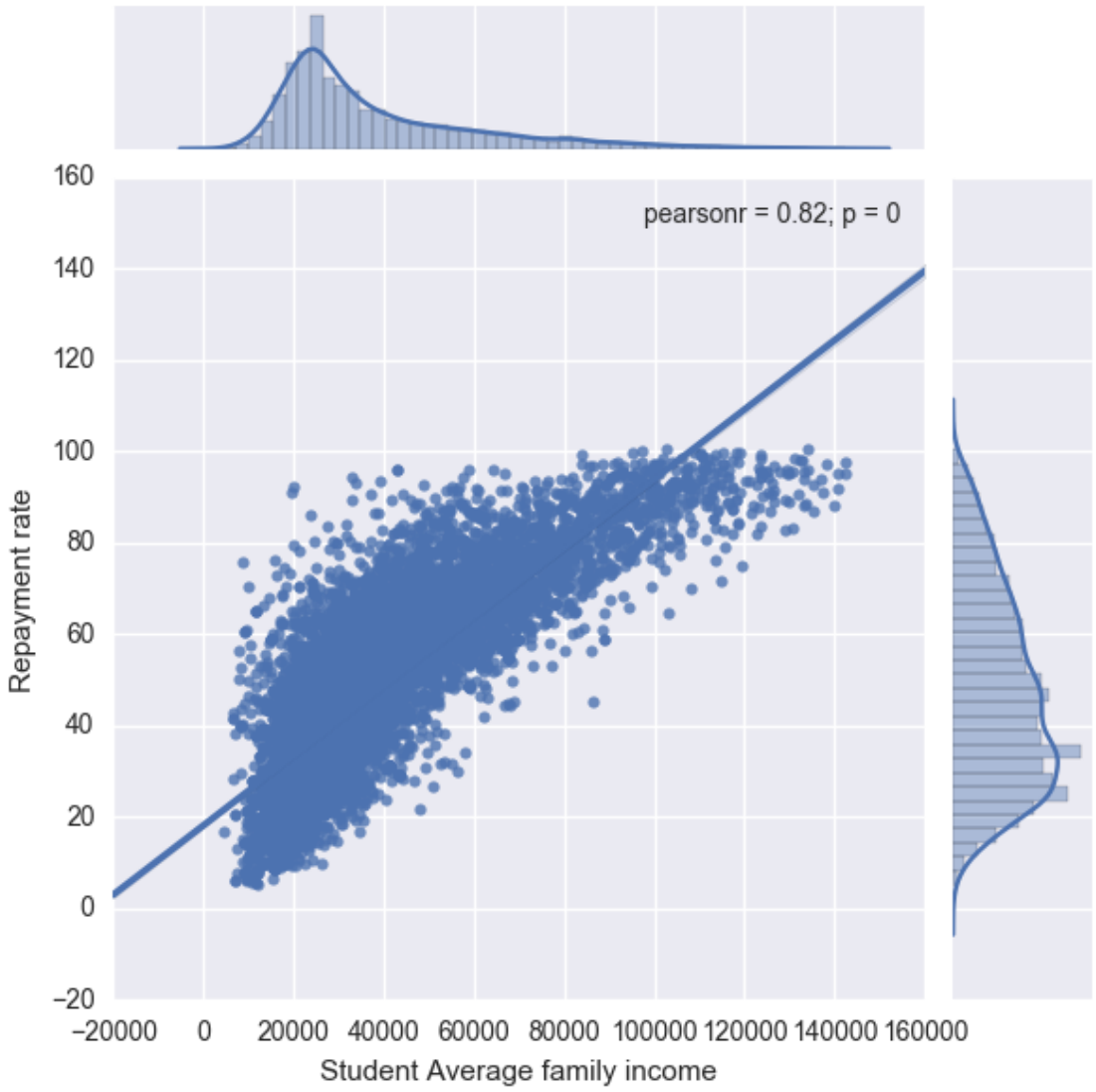
count	8705.000000
mean	47.370863
std	20.987642
min	5.162708
25%	30.228006
50%	44.855045
75%	62.622899
max	100.473631

Mean, median values are very close to each other and it seems repayment rate are normally distributed in the data. Visualisation of repayment rate can show the distribution in the data.

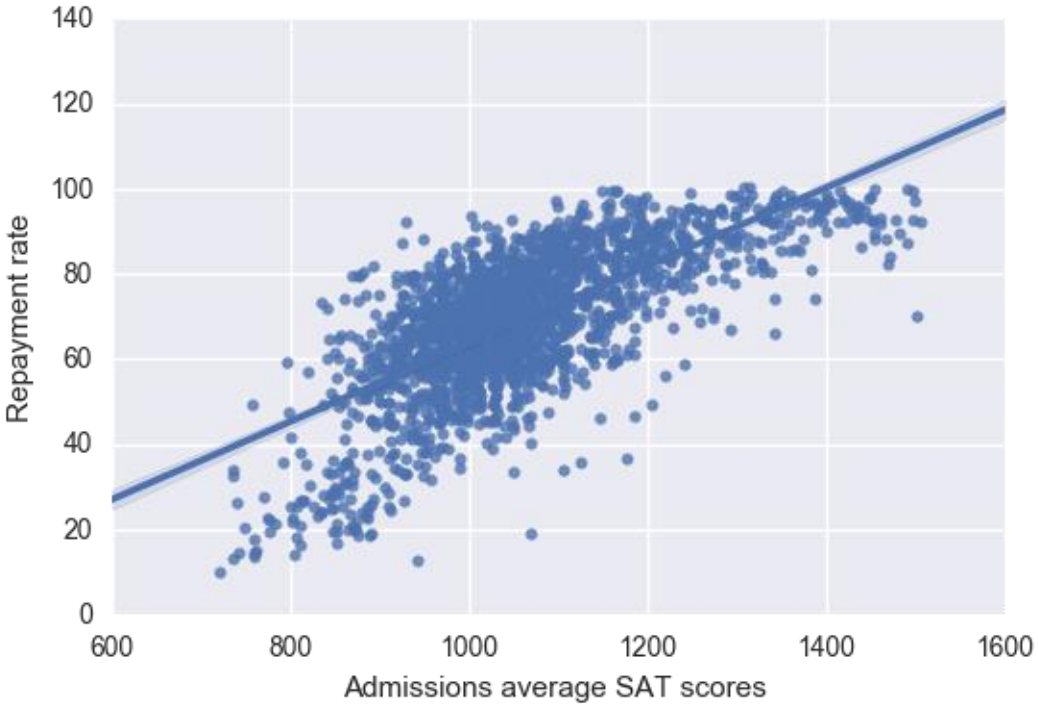


Bi-variate Analysis:

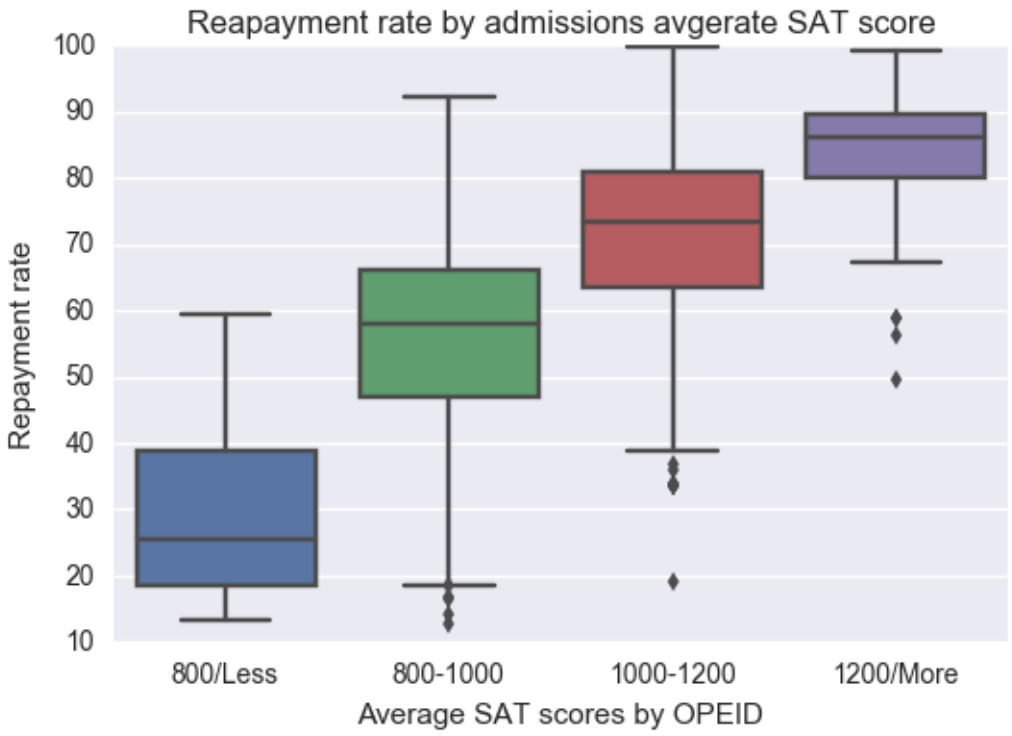
Prospective students and their families must grapple with assessing which of the many institutions available will best prepare them to achieve their goals. Cost of attendance may differ dramatically based on family income, academic background and financial aid availability. This is evident from very strong positive correlation of repayment rate with the family income.



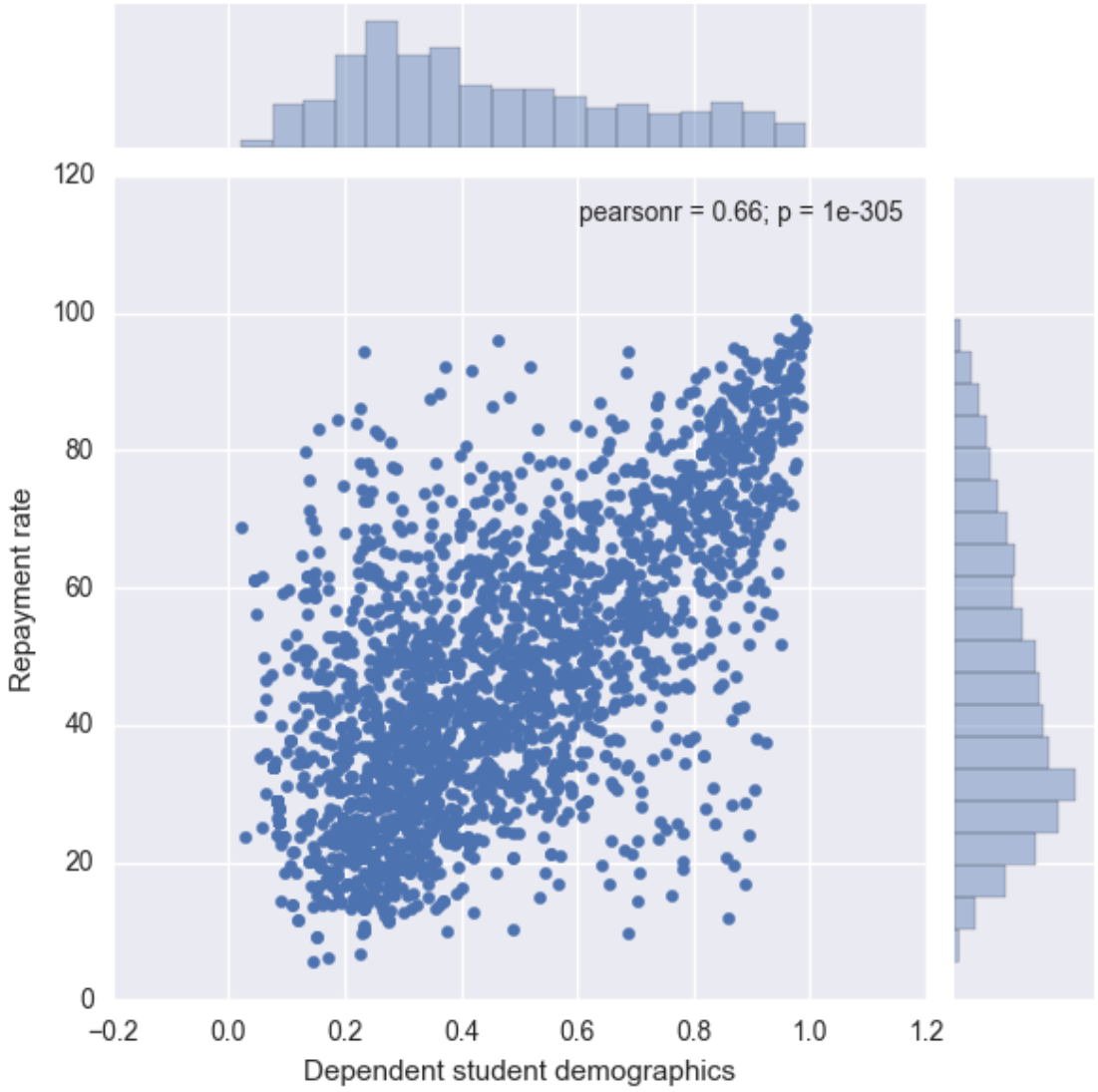
Does SAT scores determine students' ability to pay back in the future? College quality is based on comparisons of students who barely score above or below SAT test score admissions thresholds might provide a useful starting point about their relationship with repayment rate. By exploring Average SAT scores for different college, it seems there is positive relationship with repayment rate.



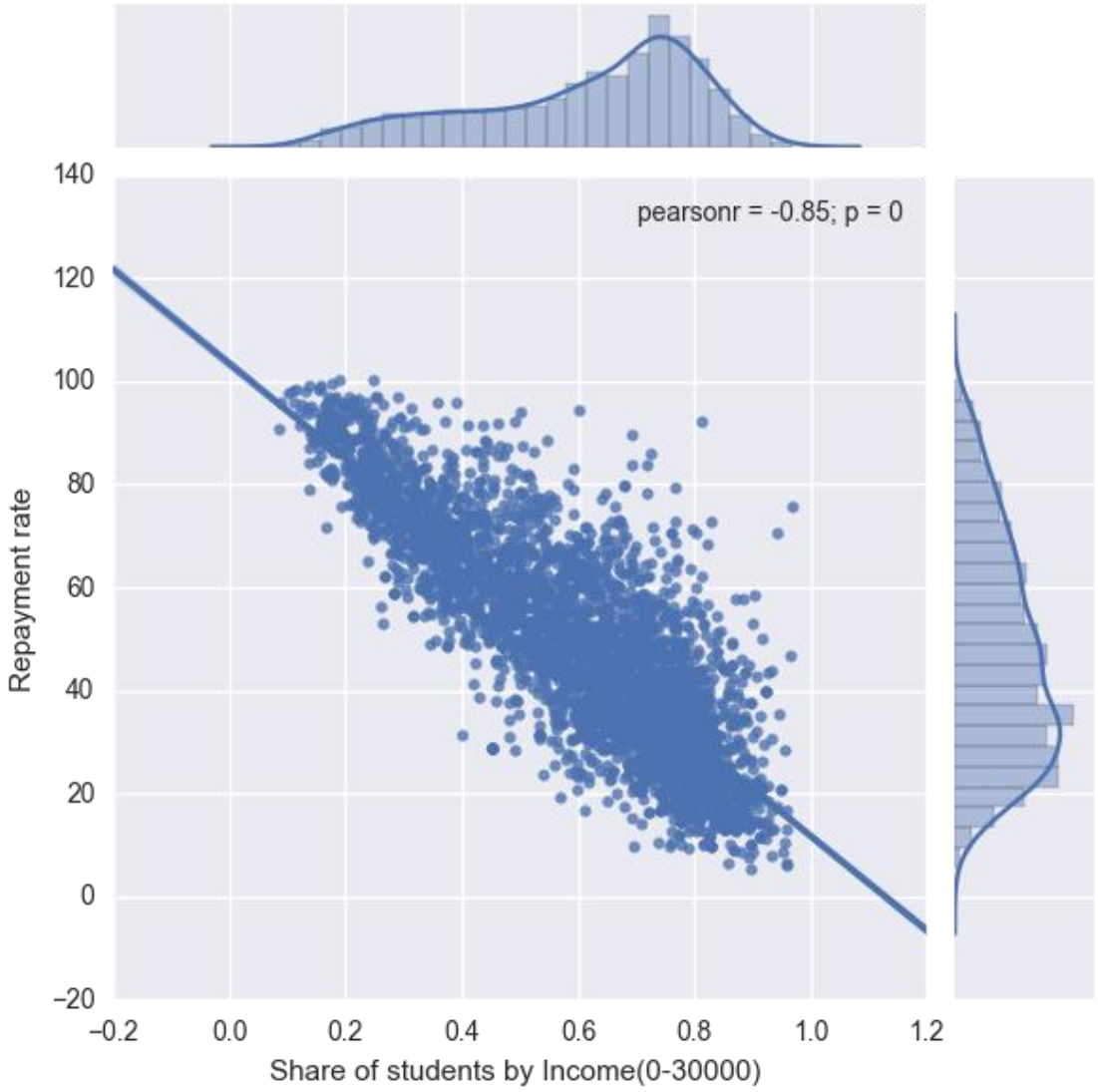
College admissions test scores (e.g., SAT or ACT scores) determine students' entry into college and academic preparation is likely to be both related to college quality. The school with highest student SAT scores have higher repayment rate than in the lower selectivity group and is likely to be evaluated more favourably than the school with the lowest (but similar) SAT score in the higher selectivity group even with identical outcomes because of the difference in peer groups.



Students from different demographic groups have different priorities. Mature students are often driven by career change or advancement goals, while younger students career goals relate to entry into a field of choice. By exploring depended students with repayment rate, this report has found student demographic is one of the major contributing factor of our predictor.



The ability to re-pay the loan is always depend on student income, which is measured by gainful employment in a recognized occupation. As such share of students in low income category has very negative correlation with the repayment rate.

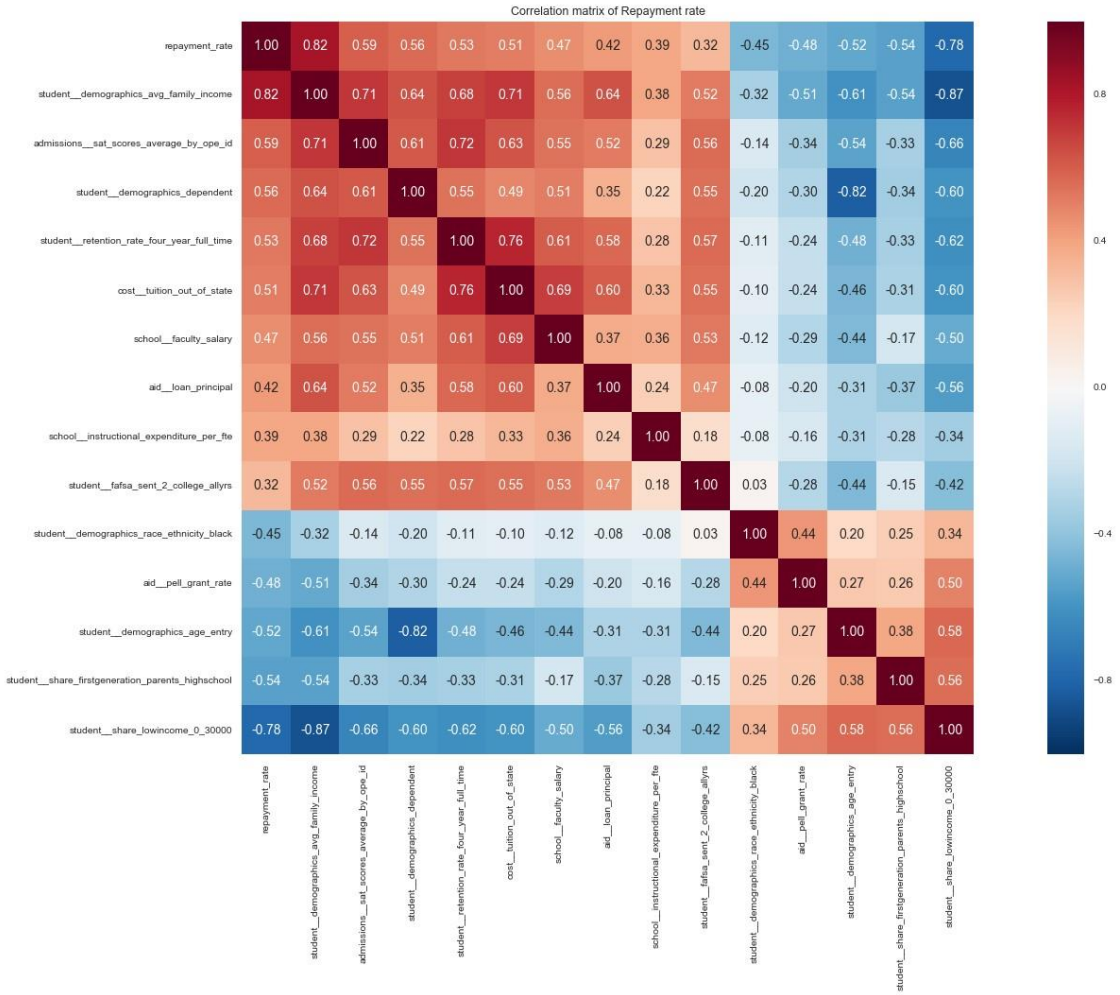


Study suggests that college costs are hard to accurately assess, especially for low-income parents who are less likely to have attended college themselves. This report finds that low-income and first-generation college students tended to overestimate the cost of college and failed to take steps in the application process that would reduce this cost, such as filling out financial aid forms which resulting low repayment rate.

Summary statistics of some of the key numeric features

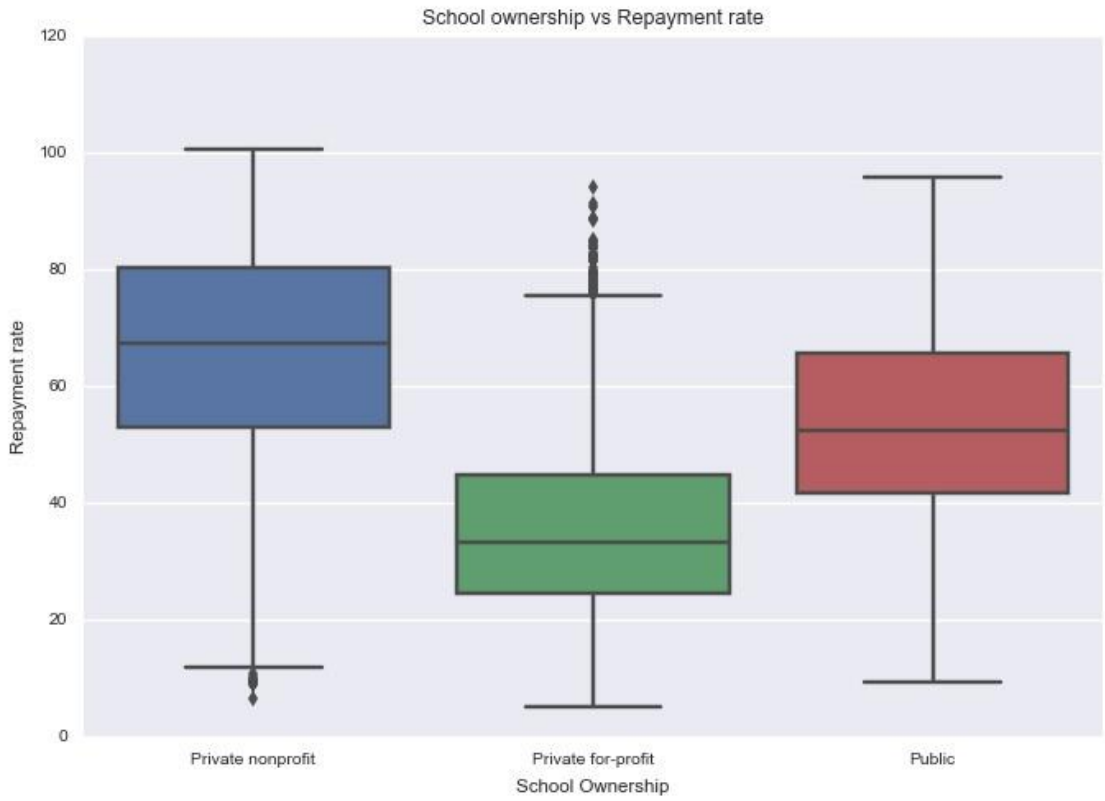
Correlation and Apparent Relationships

After exploring number of features and their relationship with our target feature, the report have found that students' family income, financial dependency status, and parents educational levels, other features have correlation with repayment rates. Following table of correlation matrix with repayment rate visualise and summarize that information.

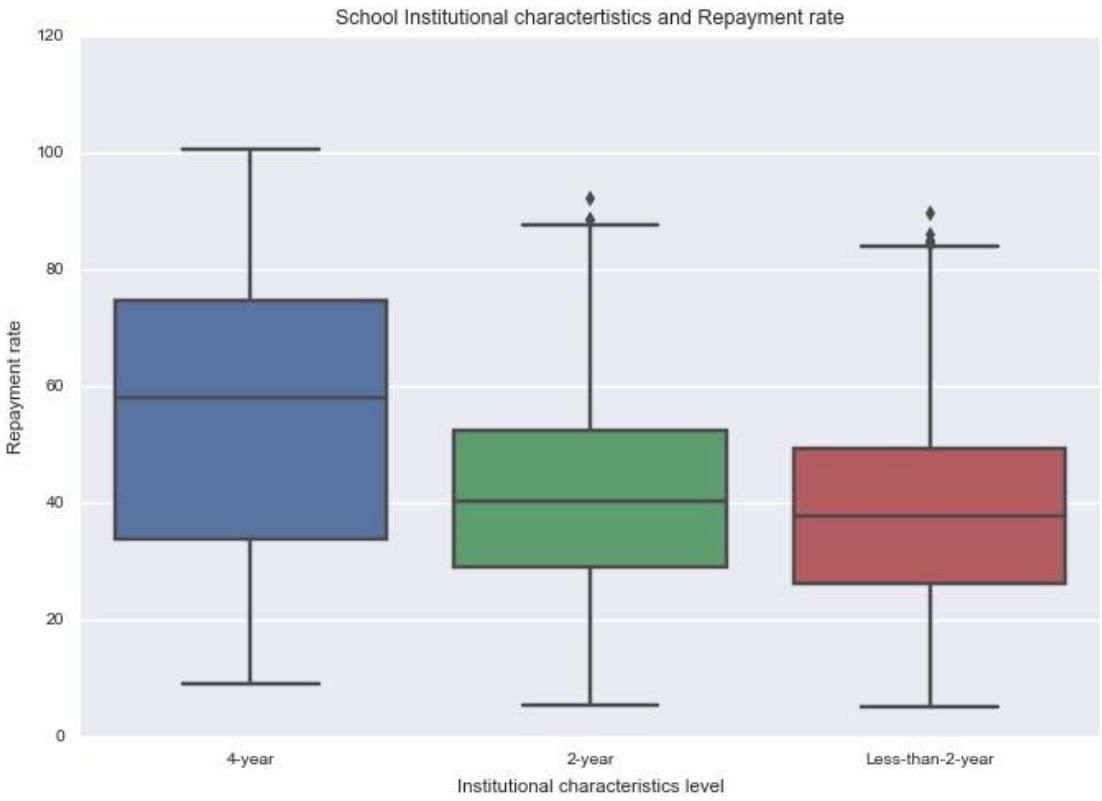


Categorical Relationships

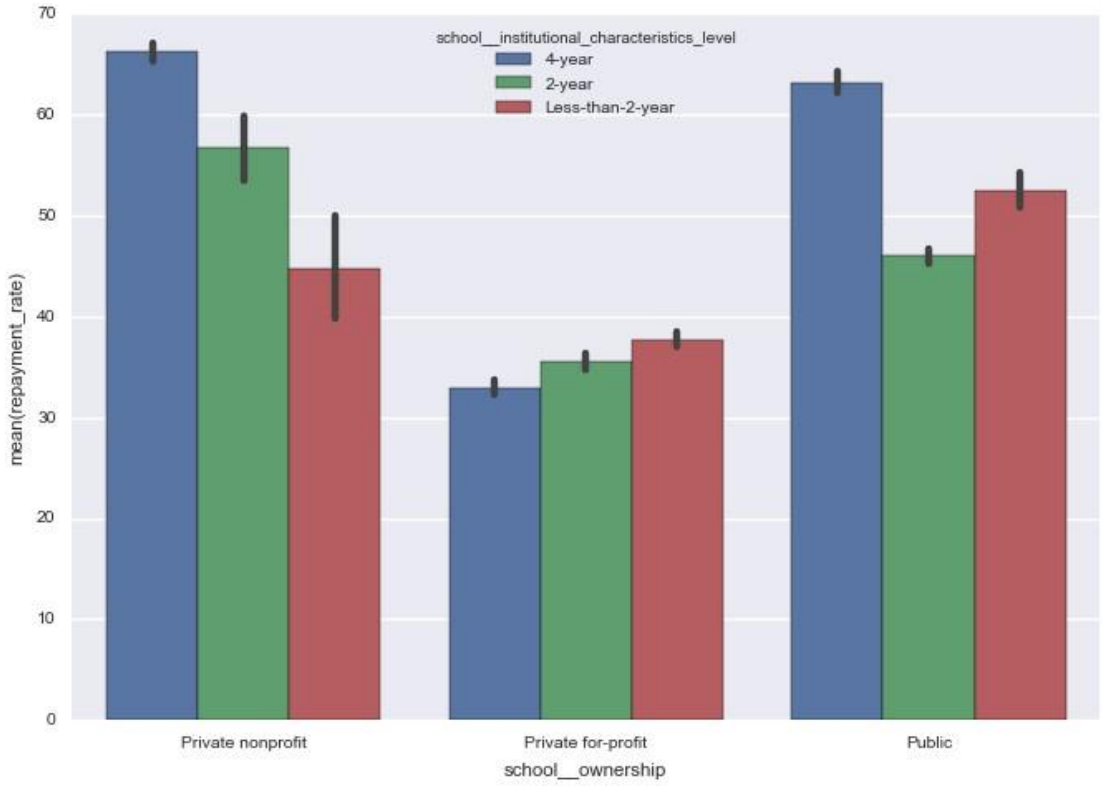
Higher education comprises a diverse range of colleges and universities that vary significantly in terms of quality and cost, this report has found, repayment rate is substantially lower in private for profit school than non-profit school.



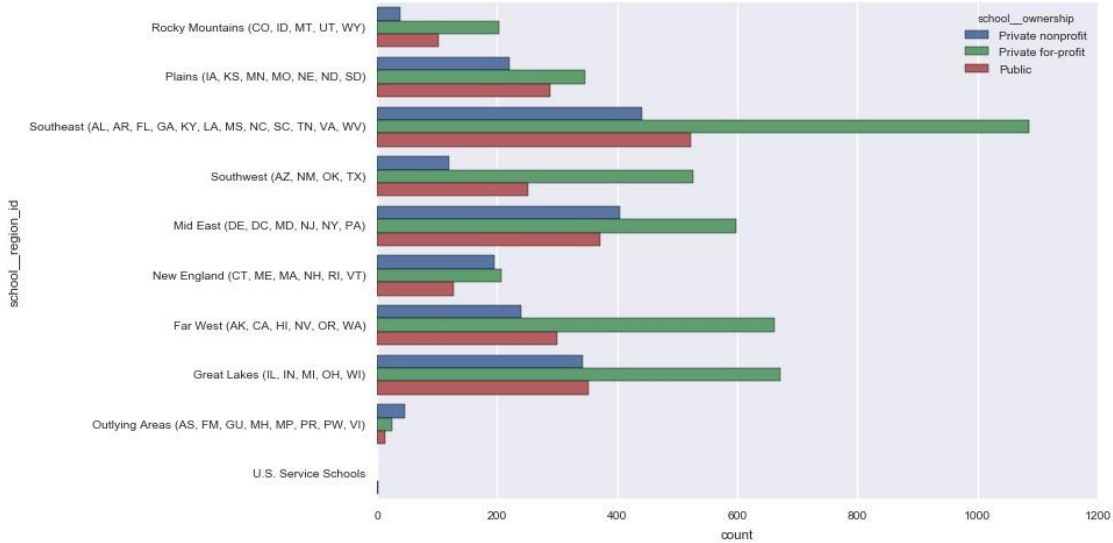
School institutions level attributable to repayment rate, while 4 years school has significantly higher repayment rates than 2 years and less than 2 years school.



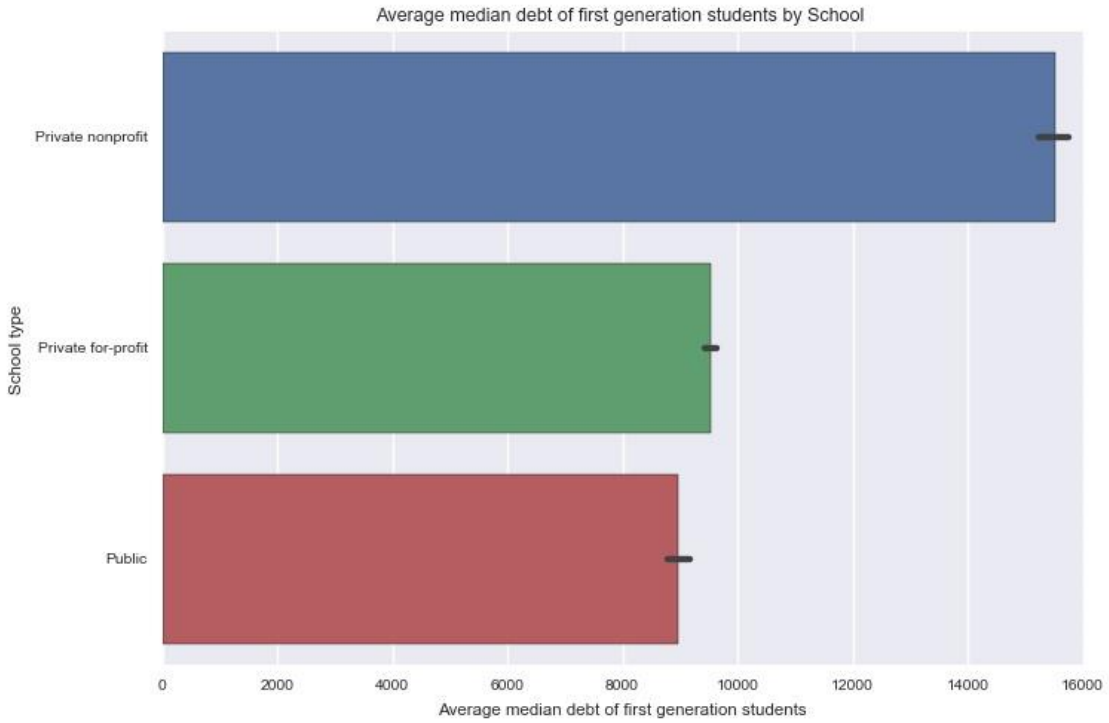
School type, the number of years to earn degree both have strong relationship with repayment rate. Repayment rate is significantly less than two-years and four-year schools in Private for profit and non-profit college.



As there are more Private for-profit schools in most regions in the datasets, so this suggests that school ownership will play key role determining the repayment rate.

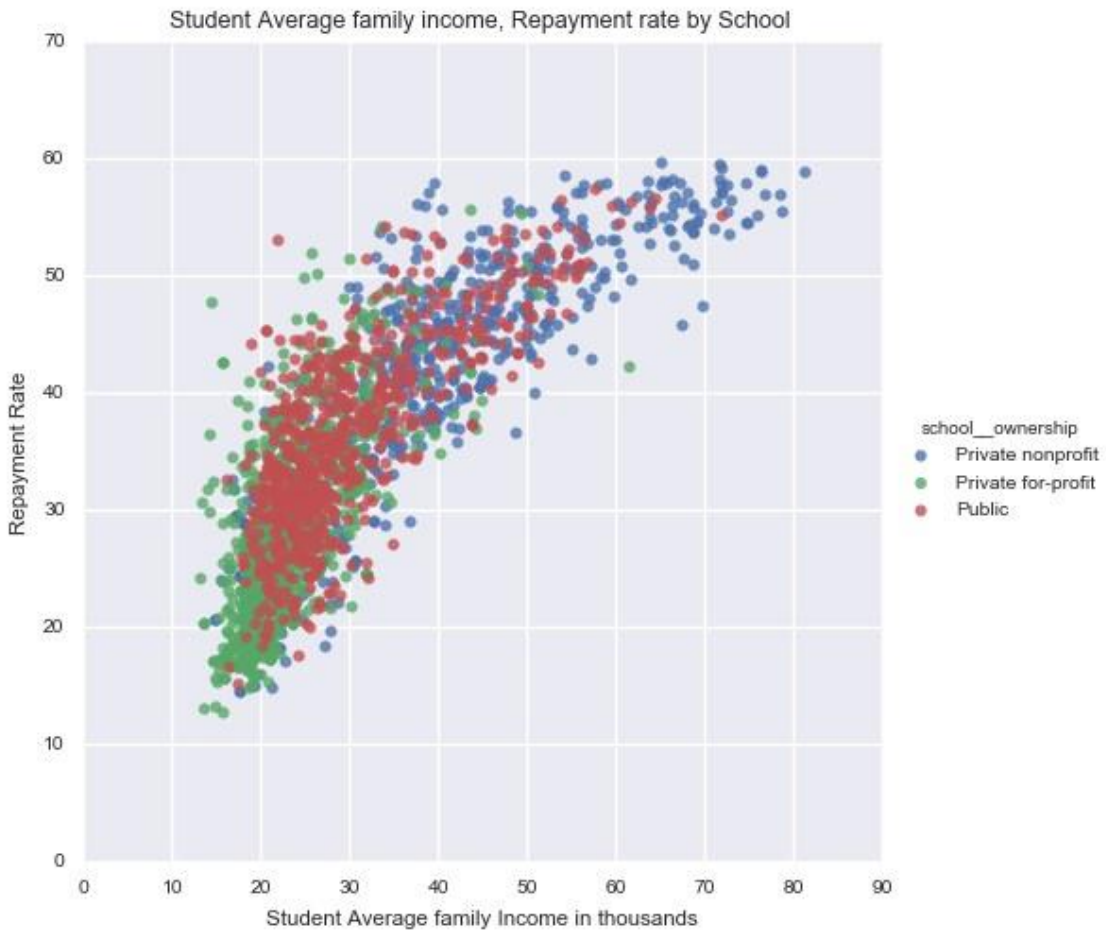


Median debt of graduates depends upon the cost of school they attended. However, relative median debt for first generation students are more in Private for non-profit school than any other schools.

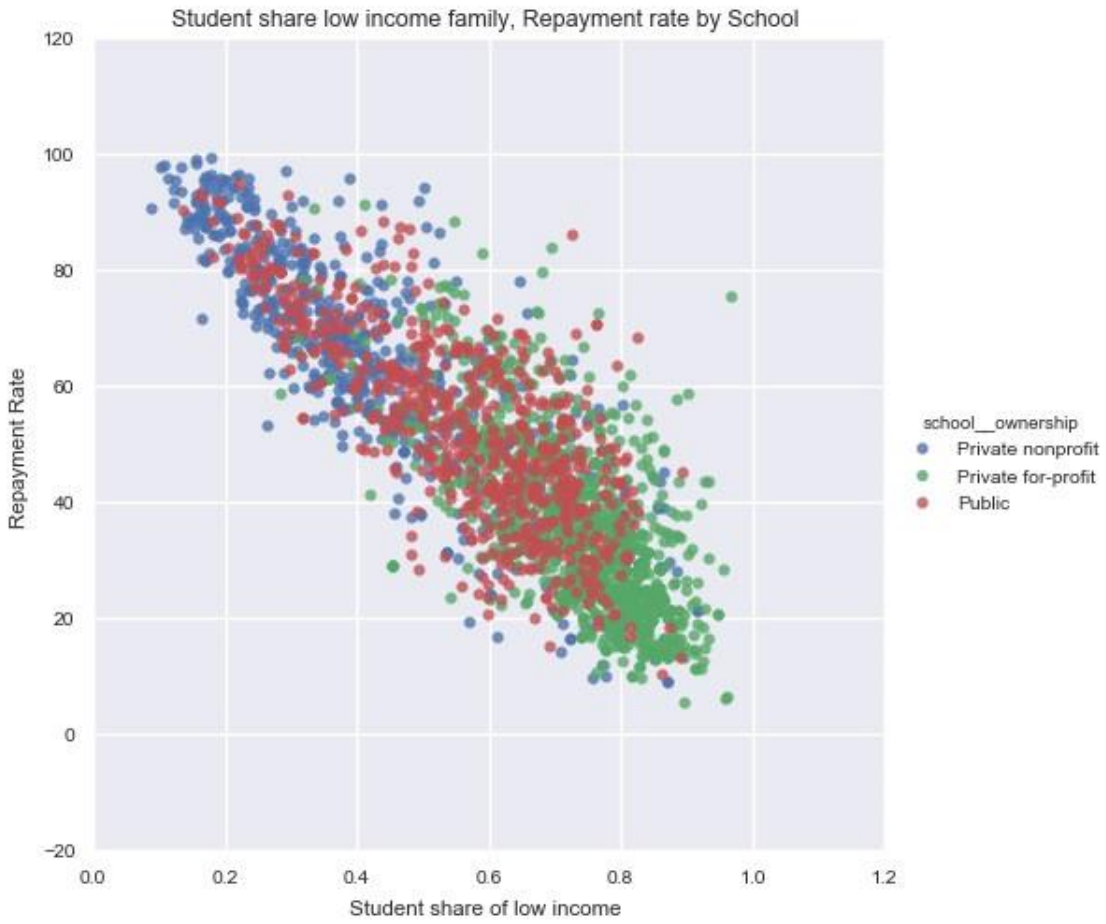


Multivariate Analysis

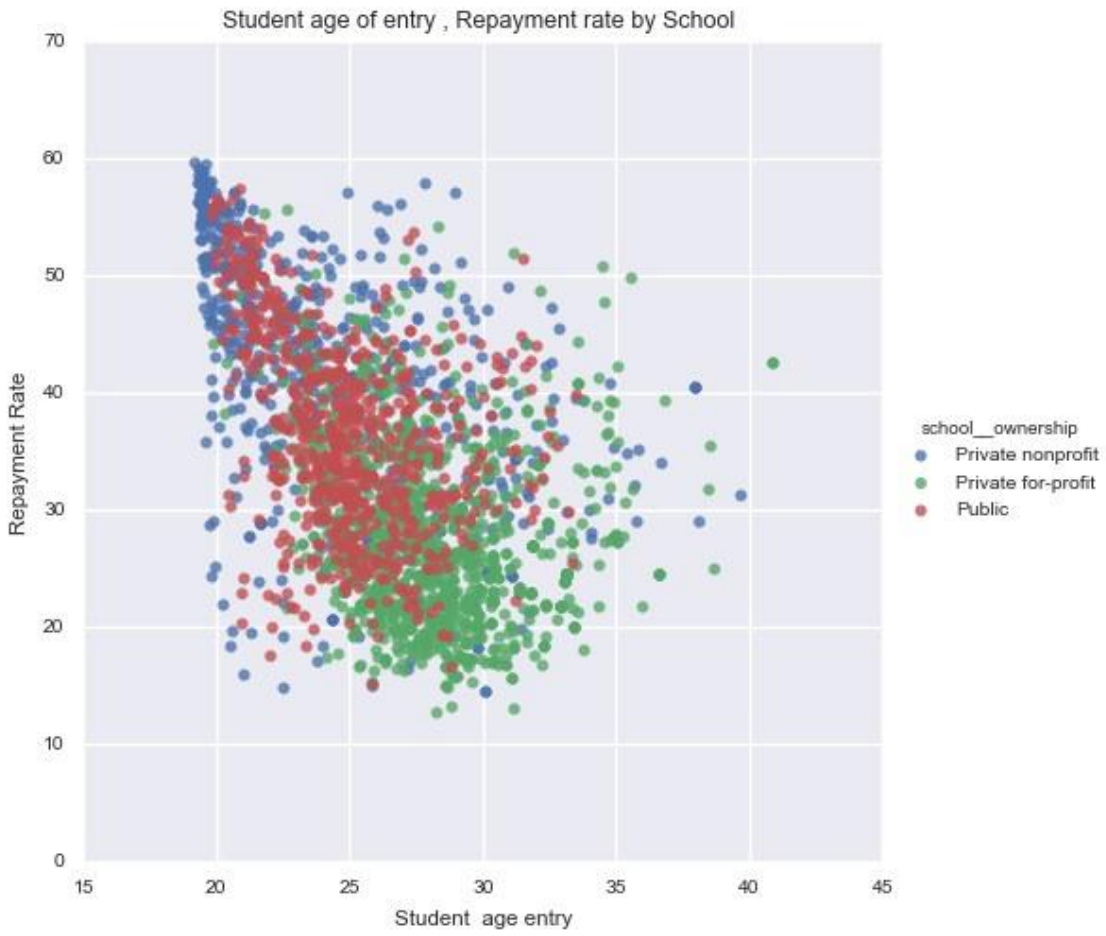
Family income has highest correlation with repayment rate. Family with low-income students in private for-profit school have least amount of repayment rate where as private for-profit school have highest repayment rate despite having relative high attendance cost.



Affording college is one of the concerns at the forefront of students and parents minds as they explore the college selection process. Families out-of-pocket costs have continued to rise, resulting higher debt. Although private for-profit college share more students coming from low income family, their repayment rate also low due to financial hardship.



Students go to college to improve employment opportunities, earn more money. Public college attendees weight affordability and location as key reasons for enrolling in a specific institution, while private, non-profit four-year students weight reputation and location most highly. Younger students have better prospect of earning in the long run as such their repayment is higher than older students. College quality also play important role in this mix.



Regression Analysis

After exploring the data regression models are developed to predict the student repayment rate based on the apparent relationships identified when analysing the data.

Linear regression models are created to measure float repayment rate value, this report uses evaluation metric called Root-mean-squared error to get the difference between actual and predicted values It is an error metric, so lower value is better to determine the performance of the model.

Models was trained with 70% of data and validated with the remaining 30% data.

Number of observations in----Training set---:6093
Number of observations in----validation set---2612
Number of values represent repayment rate ----in training data---6093
Number of values in represent actual repayment rate in ---- validation data----2612

Linear Model with L1 and L2 regularization: As seen in correlation matrix there are number of features have very strong relationship with repayment rate and those variables are correlated with each other's. Regularization is a very useful method to handle multi-collinearity, filter out noise from data and eventually prevent overfitting. Objective behind regularization is to introduce additional information(Bias) to penalize extreme parameter weights.

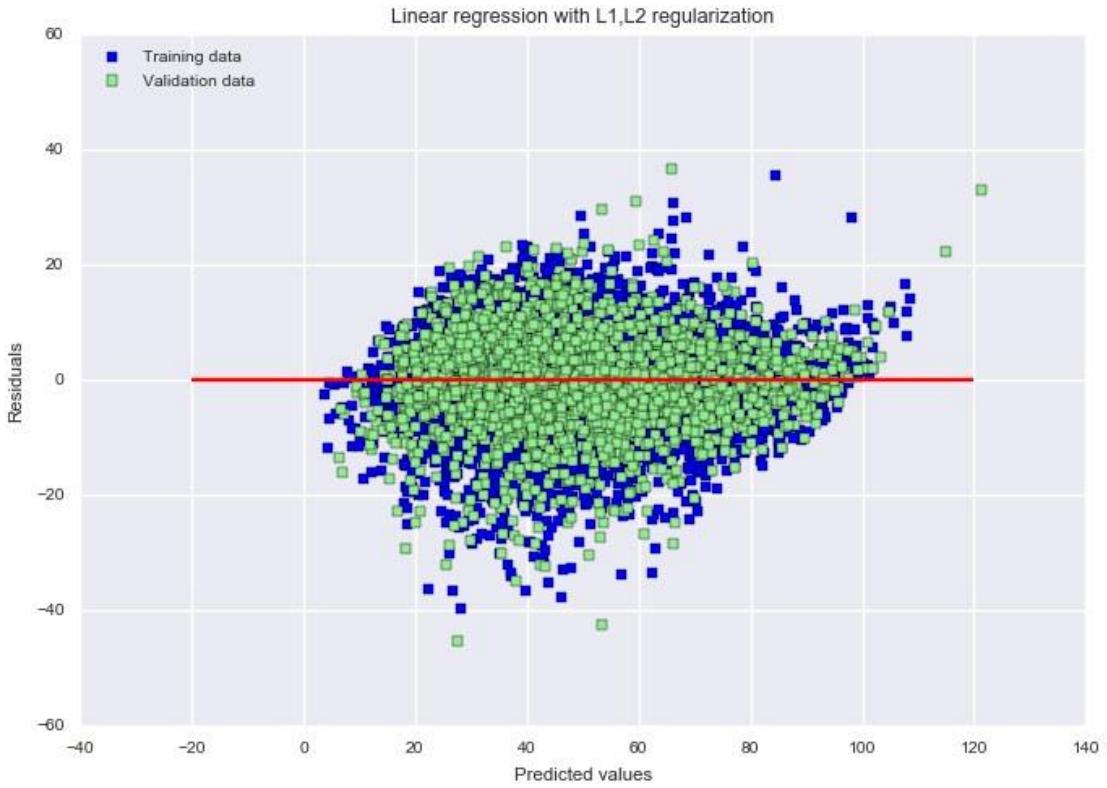
Followings are regularization parameters of the linear model.

Best l1_ratio 0.95
Best alpha: 0.03

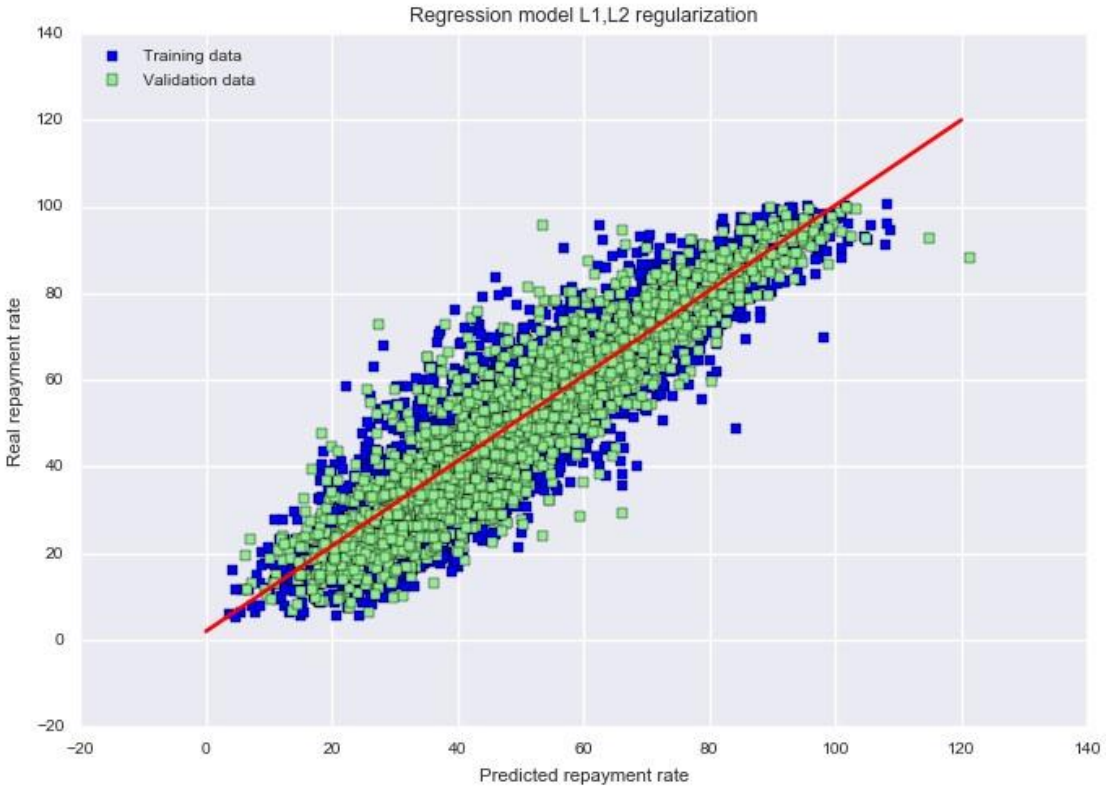
Performance Metric

Linear model with L1, L2 parameters RMSE on Training set:8.4565
Linear model with L1, L2 parameters RMSE on Validation set: 8.6432

Residual plot of the regression model



Residual plot shows it is randomly distributed and the errors are random noise A scatter plot showing the real repayment rate and actual repayment rate and the model generalize well for test data.



Goodness-of-Fit for Linear Model

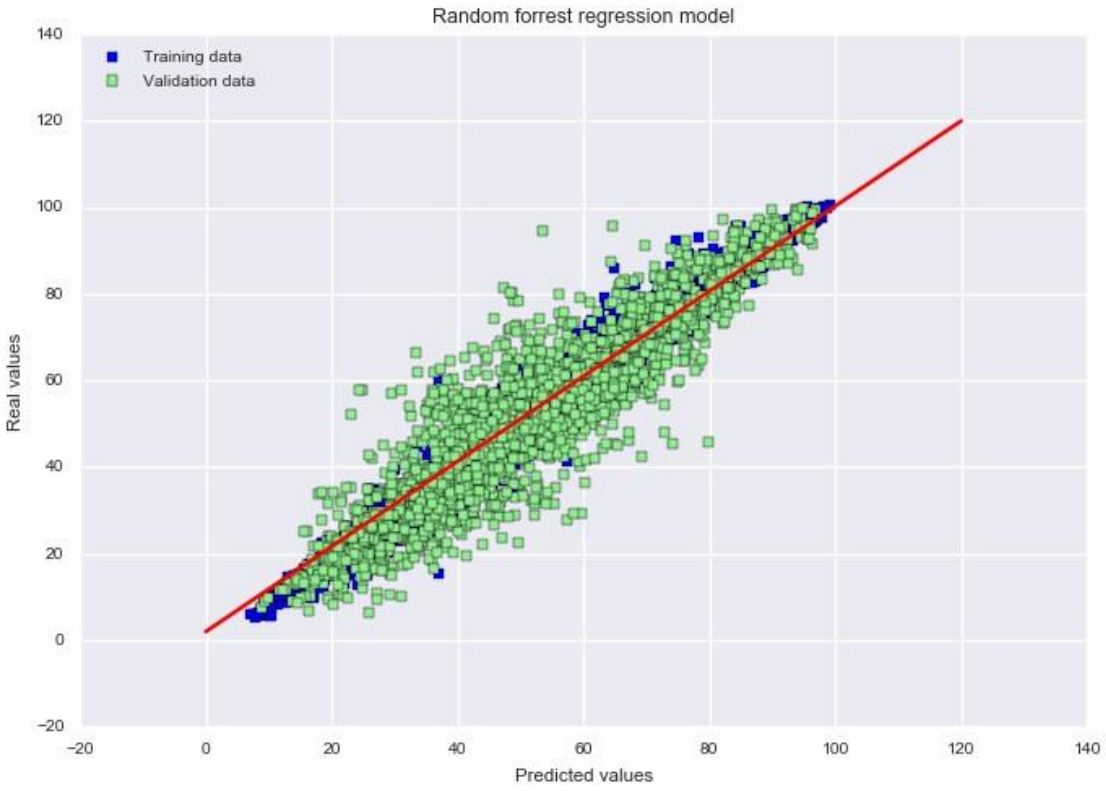
If model fits the data well the differences between the observed values and the model's predicted values are small and unbiased. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination.

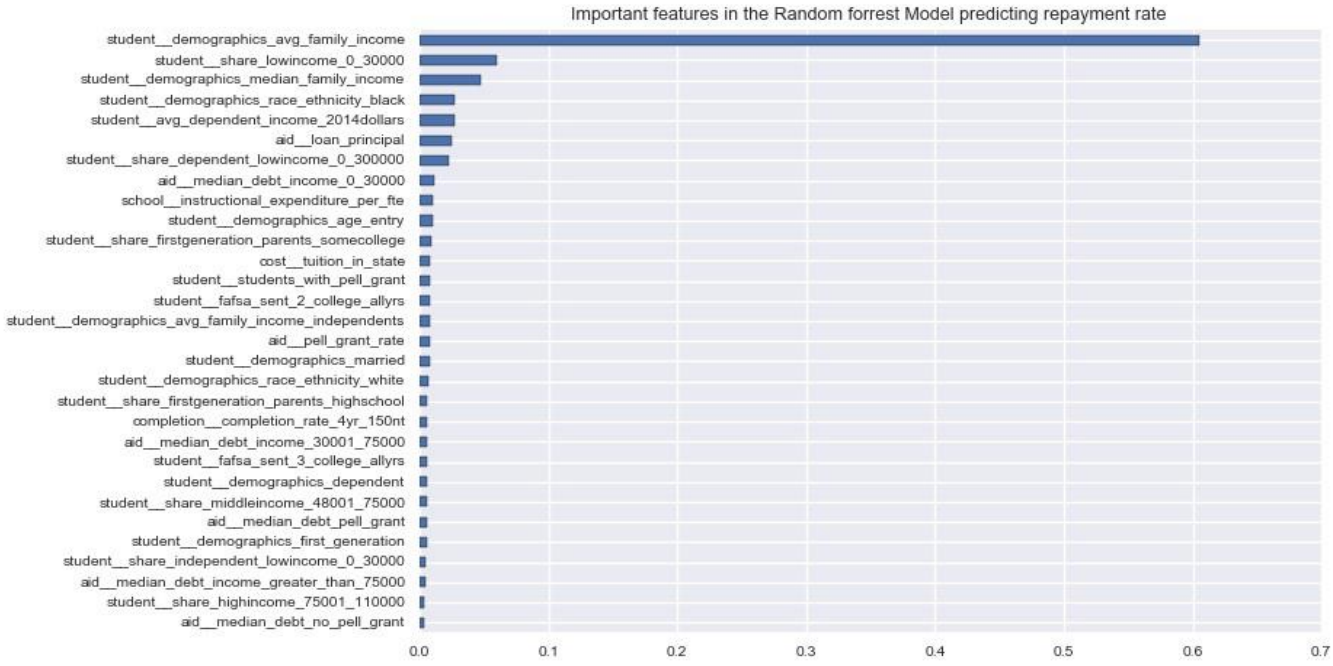
Coefficient of determination R2 of this Linear Model 0.8495

Model improvement

The Linear regression predictive model explained variation of almost 85% of the data. However, more variance that is accounted for by the regression model the closer the data points will fall to the fitted regression line. Theoretically, if a model could explain 100% of the variance, the fitted values would always equal the observed values and, therefore, all the data points would fall on the fitted regression line. For improvement, a Random forest regression model is developed with the training data and better performance is observed of RMSE and R2 value.

RMSE value of training data 7.6241
Coefficient of determination R2 of Random forest regression model 0.9210





Conclusion

The analysis shows student repayment rate can be predicted, given the fraction of lower-income students that institutions enrol, including those defined by family income, first-generation status, age, ethnicity as well as cost of attendance, price of tuition, students' debt, any aid assistance like Pell grant. College performance is key measures which drive future earnings, debt management finally repayment rate. All students should have access to basic information about the financial consequences of their decisions so they can pursue their passions with a clear understanding of the economic trade-offs.