Predicting the Student Loan Repayment

Project Report by
Jawad Rasool

July 2017

1. Executive Summary

In this document, the author presents his findings from the analysis of the data related to loans given to students at United States institutions of higher education, and provides a description of the model to predict the repayment rate for the student loans.

Student loan is a complex data issue. A student's ability to repay loans depends on the job they get after graduation. This in turn depends on the school they went to, the degree they earned, the area they live in, their family circumstances, and how much money they borrowed. By modeling what percent of students repay their loans, an attempt is made to better understand the properties of schools that make them better or worst investments, in terms of whether attending that school increases the probability that the students will get a job good enough to repay their loans.

This analysis is based on 8705 observations (identified by a unique identifier "row_id") and 443 features/variables in this dataset (excluding row_id and repayment_rate). Out of these, 213 of them are identified as categorical features and rest are numeric features. Each row in the dataset represents a United States institution of higher education and a specific year (i.e. two particular years, denoted by "year_a" and "year_b"). Our goal is to predict the variable "**Repayment Rate**". It is defined as the approximate percent of students that make active repayments on their loans. For example, a value 25 means that about 25% of the students have been decreasing the balance on their loans through repayment.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several relationships between the features and the Repayment Rate are identified. We then create a model to classify the USA institutions into two classes based on whether the repayment rate is lower or higher than the average repayment rate. Finally a regression model to predict the Repayment Rate from its features is created. The performance metric considered in this regression problem is Root Mean Squared Error (RMSE).

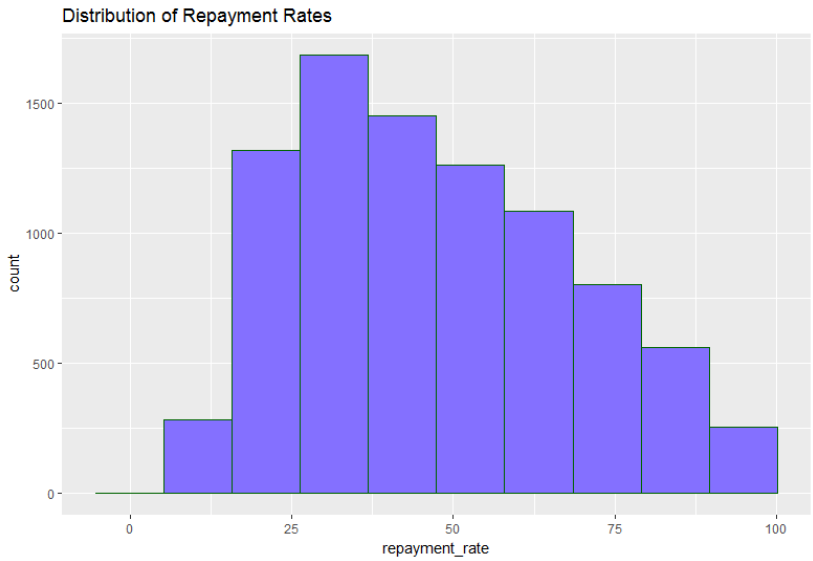
2. Data Exploration

2.1. Repayment Rate

We first look at the Repayment Rate (given by the variable "repayment_rate"). Below are the summary/descriptive statistics of this variable:

Minimum	Maximum	Median	Mean	Std. Dev
5.163	100.474	44.855	47.371	20.988

The statistics above shows that the mean and the median values are not significantly different. But since the median is lower than the mean, the repayment_rate values are slightly right-skewed. A histogram of the repayment_rate variable (shown below) also confirms the slight right-skewness:

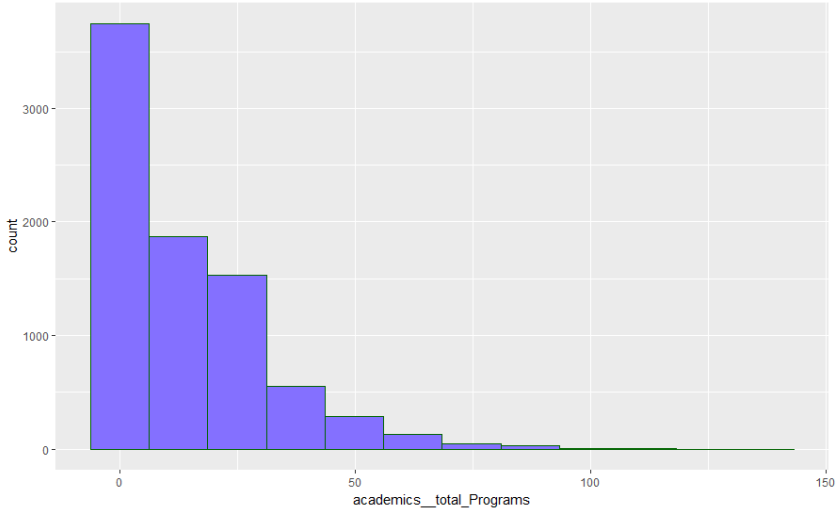


2.2. Feature Engineering

As mentioned earlier, there are 213 categorical features in the dataset. We first created a new feature "**academic_total_programs_offered**" indicating the total number of programs offered at each school. The summary statistics for this new variable are provided below:

Minimum	Maximum	Median	Mean	Std. Dev
0.00	137.00	9.00	14.55	16.077

The histogram shows that most of the schools offer less than fifteen programs. We also see that the histogram is right-skewed:



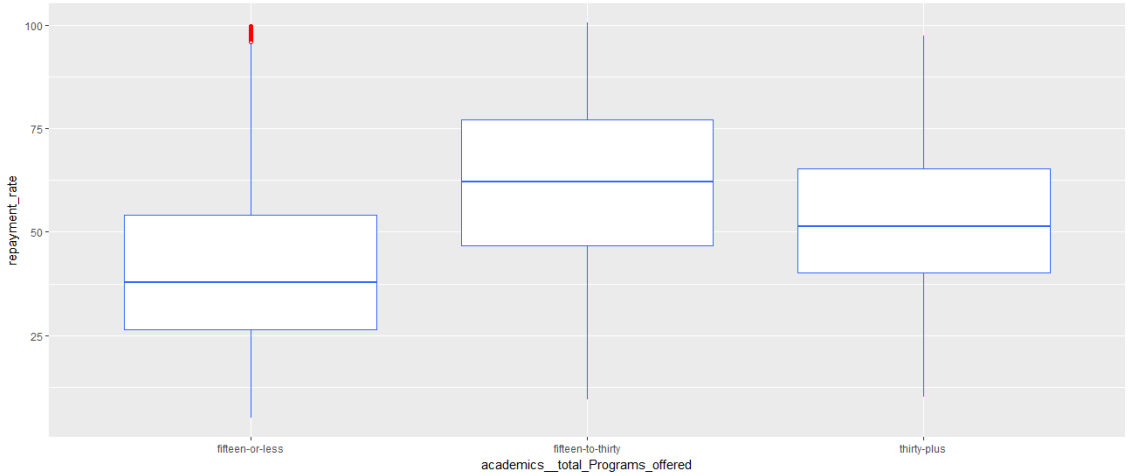
We then created a categorical feature based on the above variable, consisting of three values:

- fifteen-or-less
- sixteen-thirty
- thirty-plus.

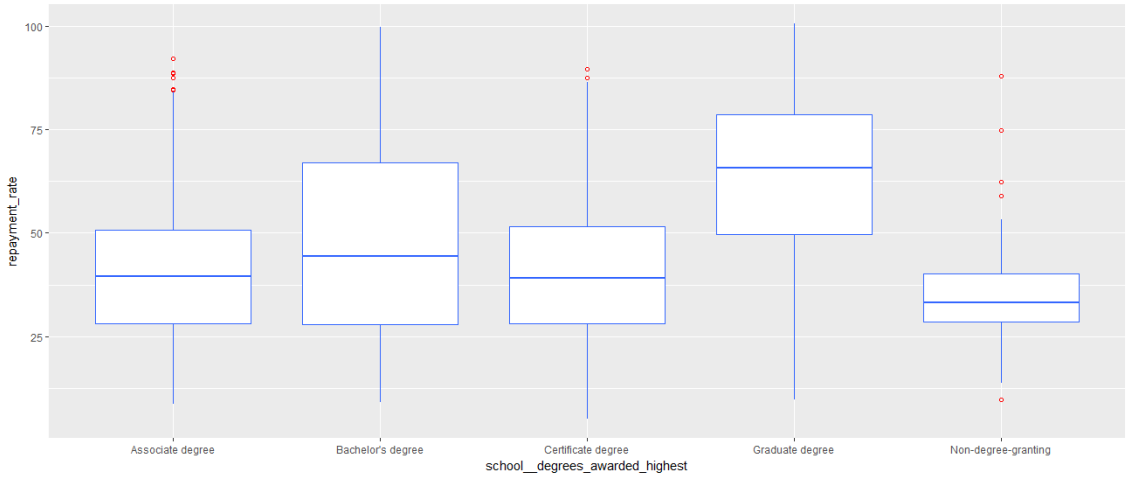
2.3. Categorical Relationships

The author then proceeded with the exploration of the relationship between categorical features and Repayment Rate. The following box plots show some of the categorical features that seem to exhibit a relationship with the Repayment Rate:

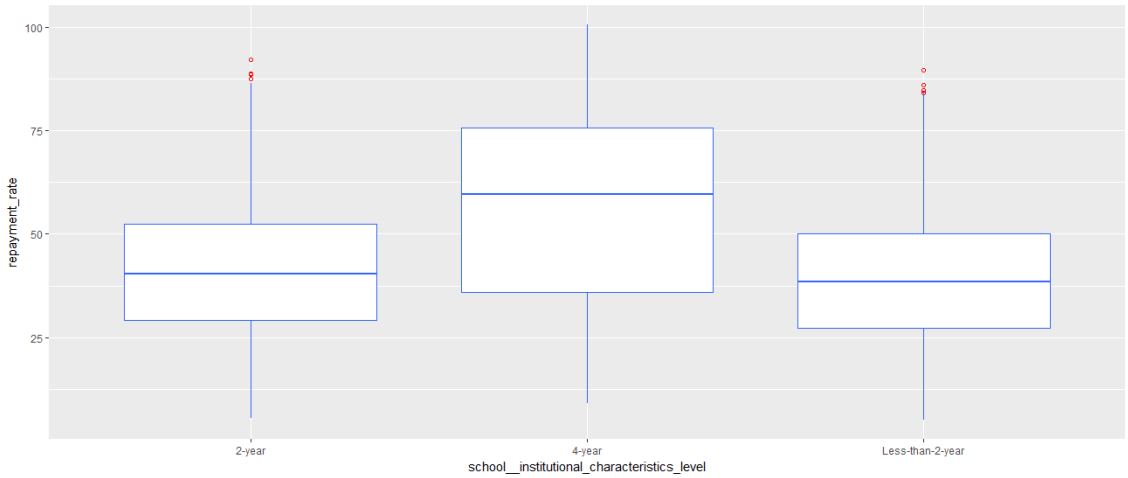
- The median repayment rate for institutions offering less than fifteen programs is the lowest:



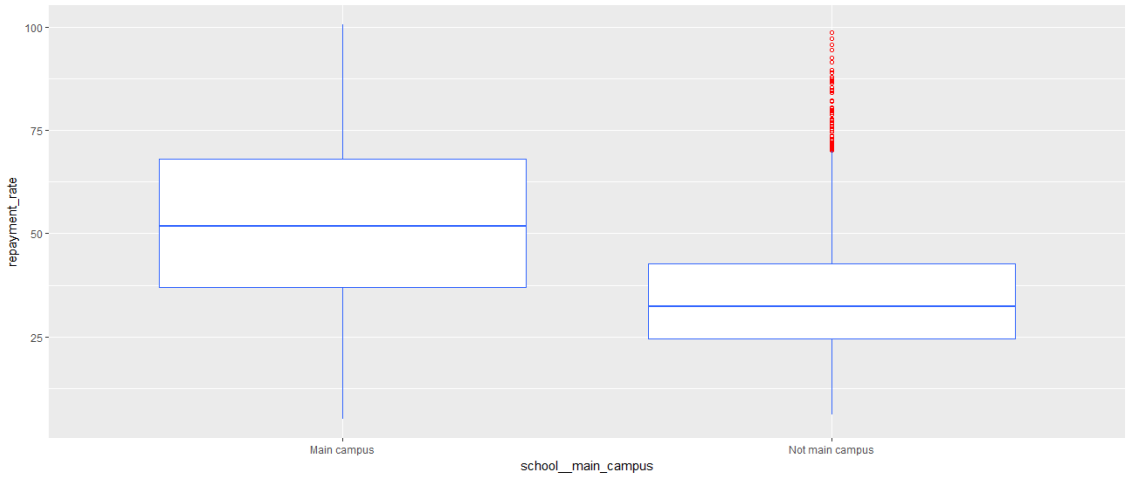
- The institutions which award the Graduation degree as their highest degree have the highest median repayment rate. Non-degree granting schools have the lowest median repayment rate:



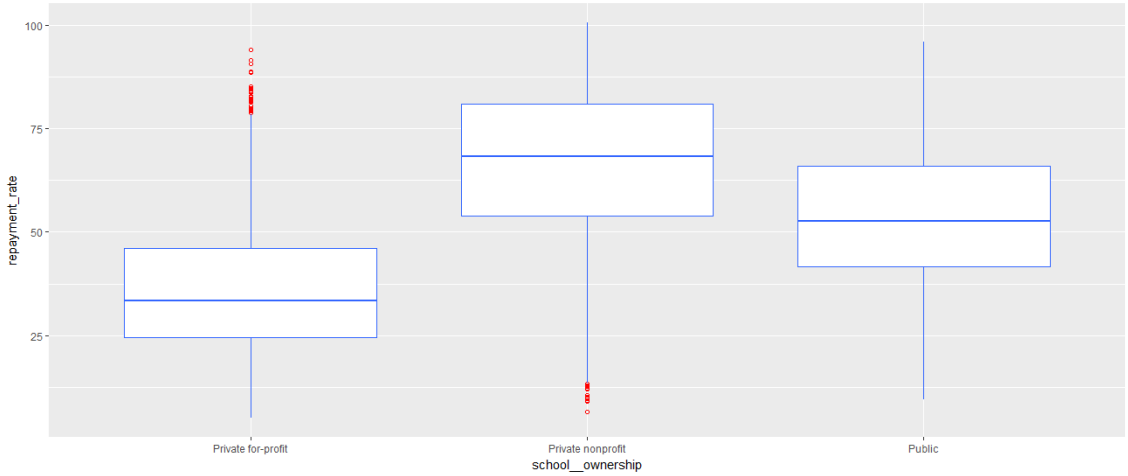
- The institutions with “4-year” programs have the highest median repayment rate. They also seem to have the highest range of values for the repayment rate:



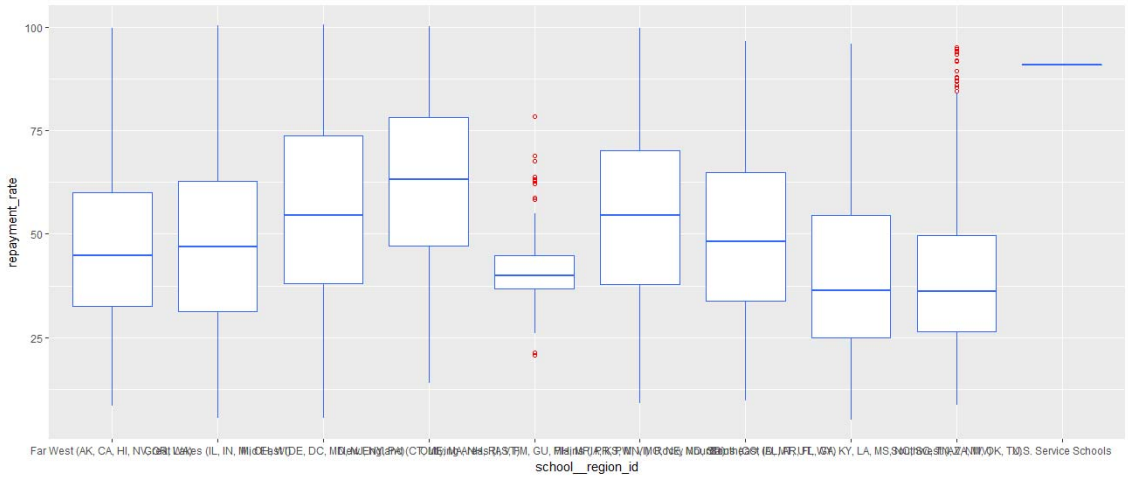
- The students studying at the main campus seems to have a higher median repayment rate:



- The median repayment rate for privately owned not-for-profit schools is higher than that of public schools. The median repayment rate for privately owned for-profit schools is the lowest:

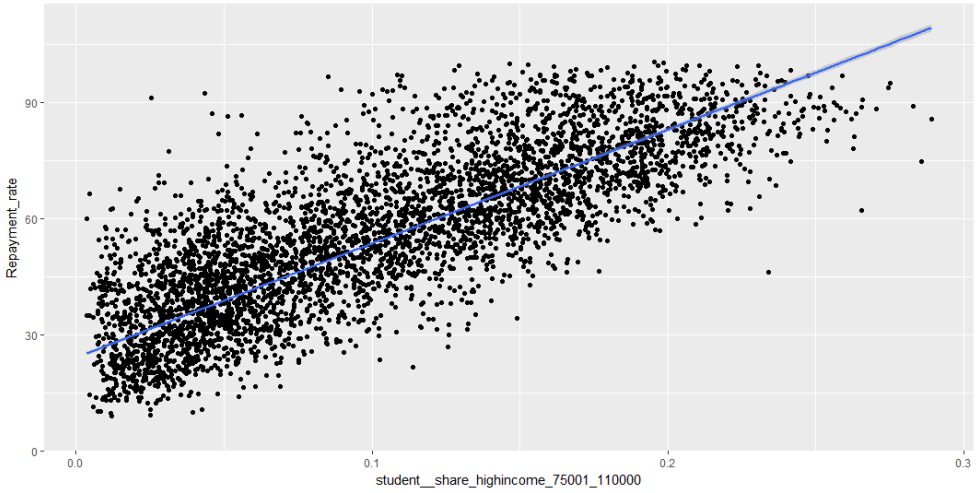


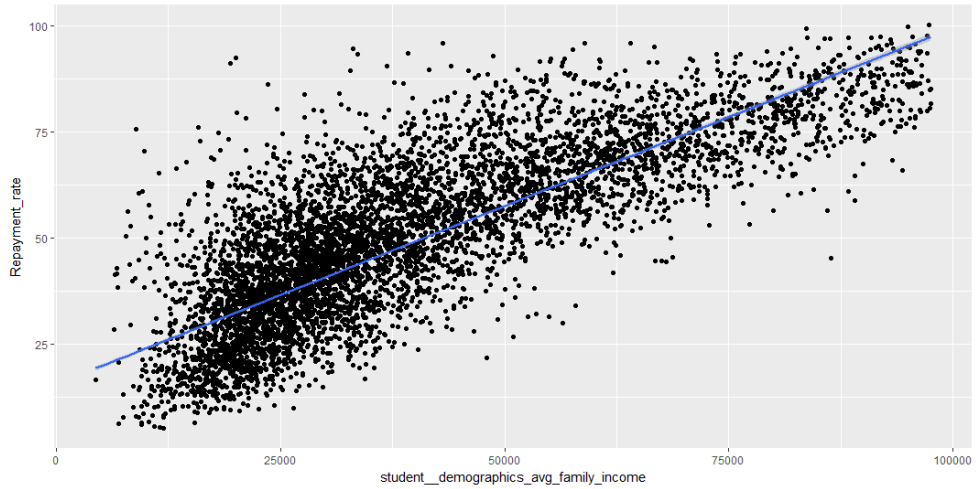
- Among various school regions, New England has the highest median repayment rate:

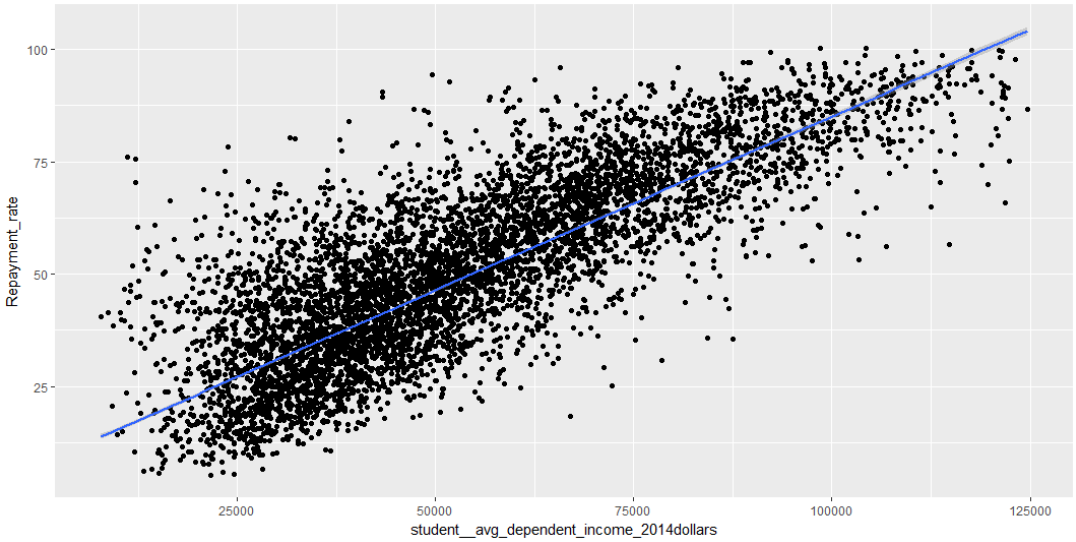


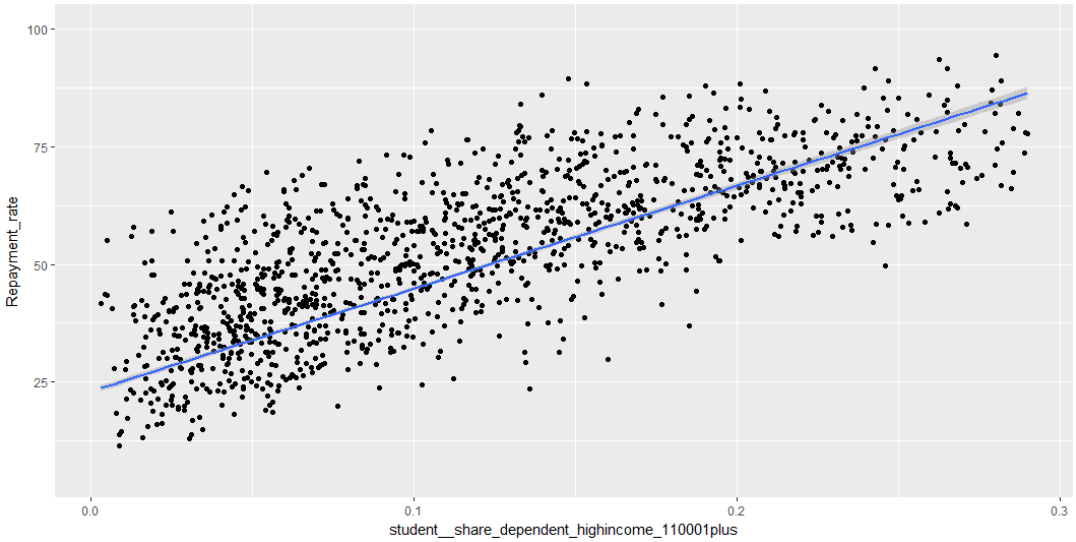
2.4. Numeric Relationships

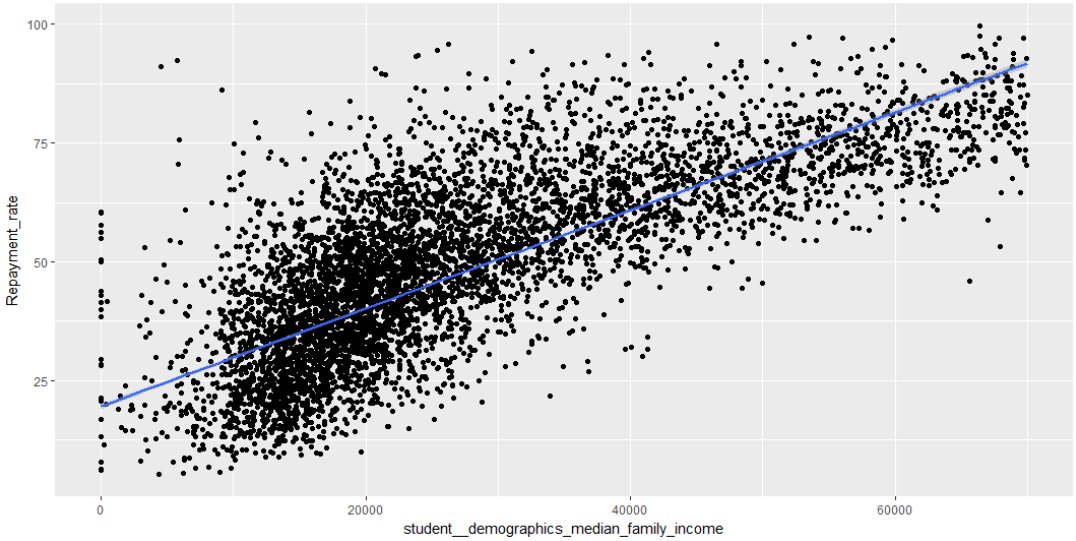
Since there are more than 300 numeric features, it is very difficult to present the summary statistics for all of them. The author has, therefore, looked at the scatter plots for these numeric features against the repayment rate, and plotted some of them below:

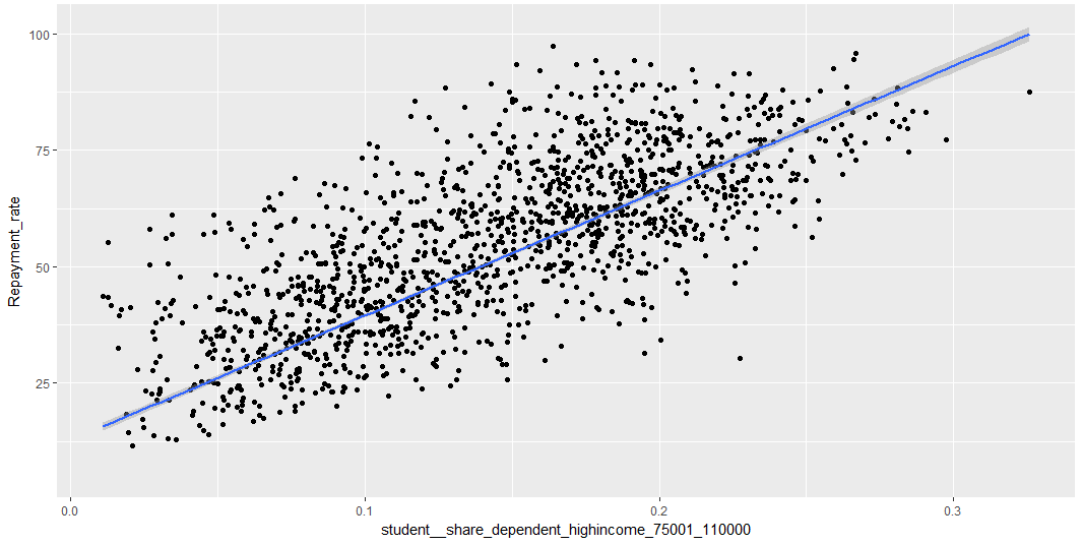


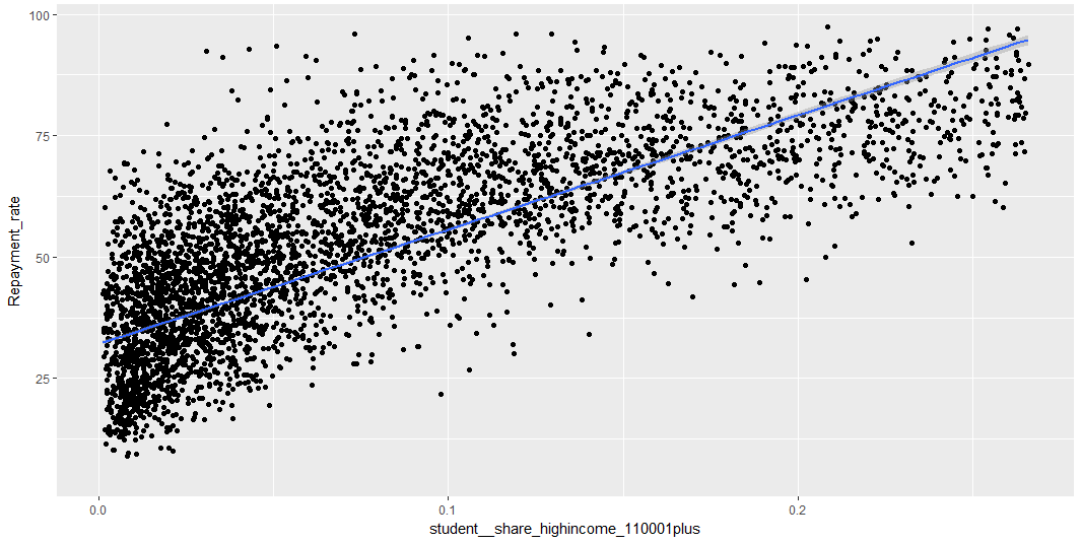


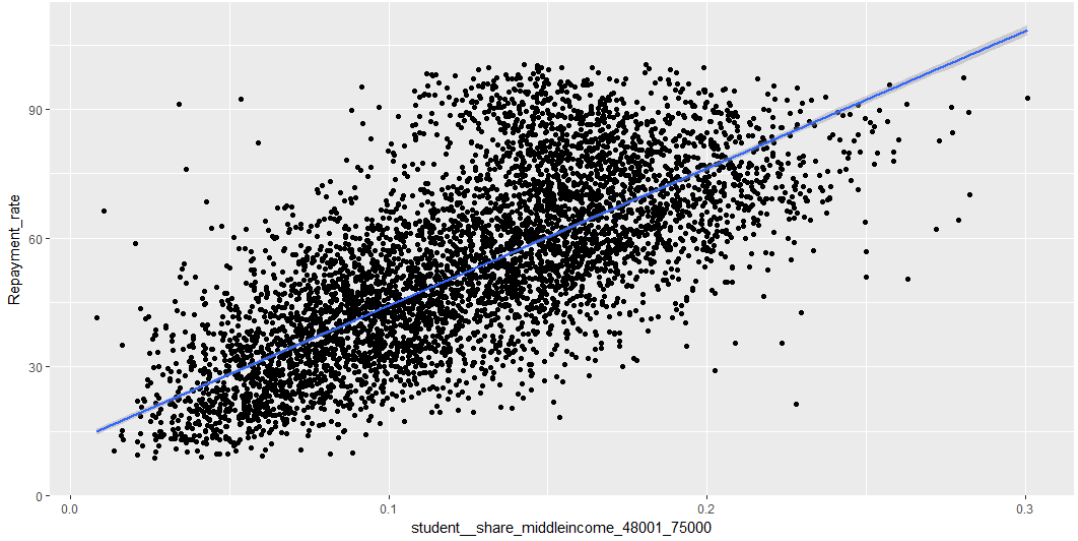

Repayment_rate
100
75
50
25
25000
50000
75000
100000
125000
student__avg_dependent_income_2014dollars

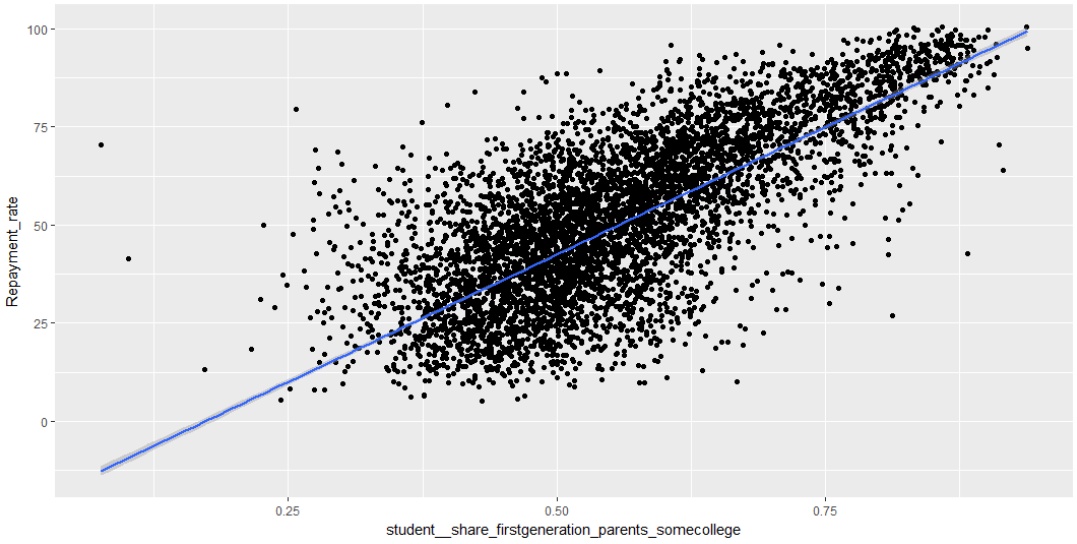

Repayment_rate
100
75
50
25
0.0
0.1
0.2
0.3
student__share_dependent_highincome_110001plus

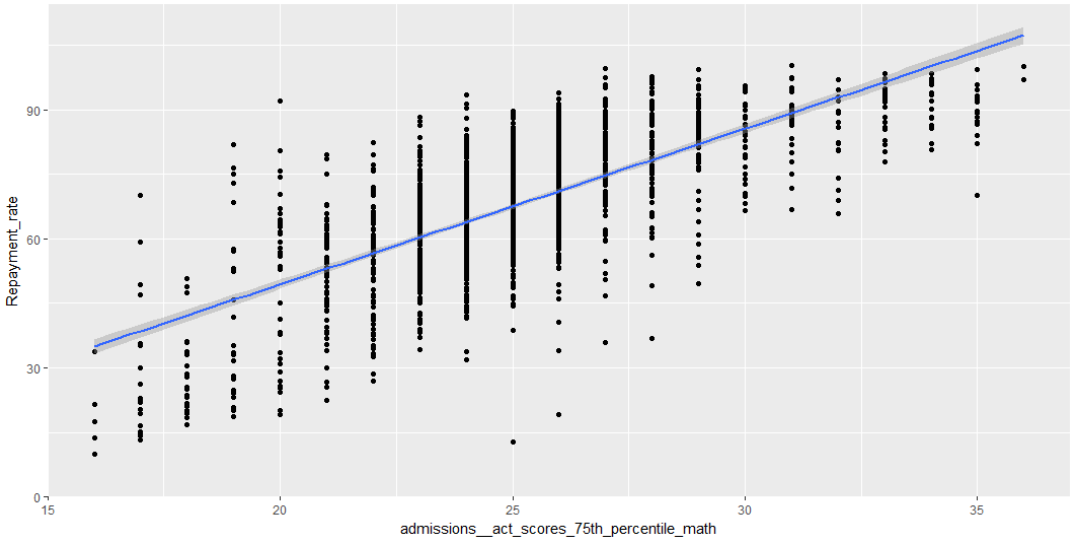

Repayment_rate
100
75
50
25
0
20000
40000
60000
student__demographics_median_family_income

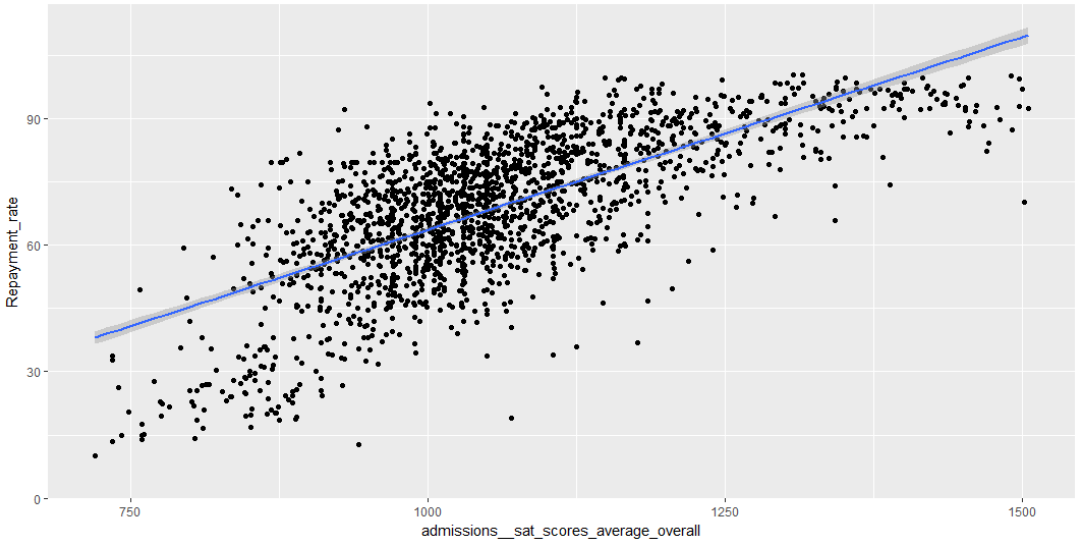

Repayment_rate
100
75
50
25
0.0
0.1
0.2
0.3
student__share_dependent_highincome_75001_110000

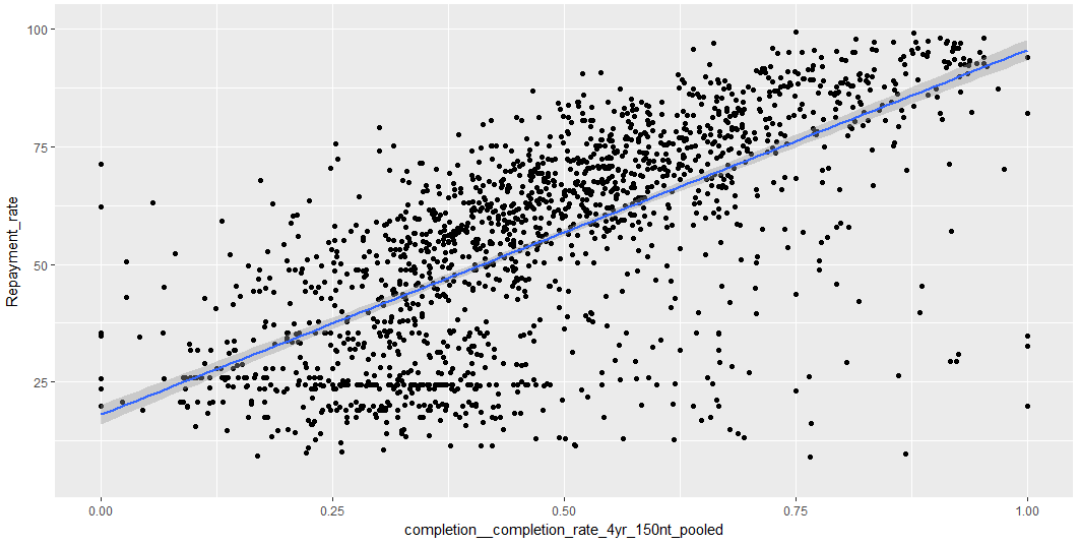

Repayment_rate
100
75
50
25
0.0
0.1
0.2
student__share_highincome_110001plus

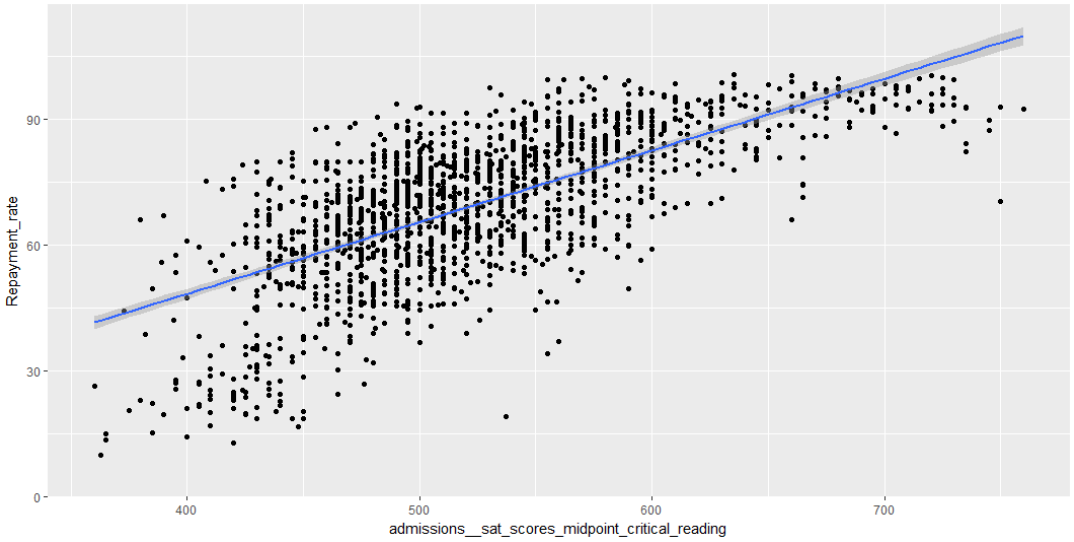

Repayment_rate
90
60
30
0
0.0
0.1
0.2
0.3
student__share_middleincome_48001_75000

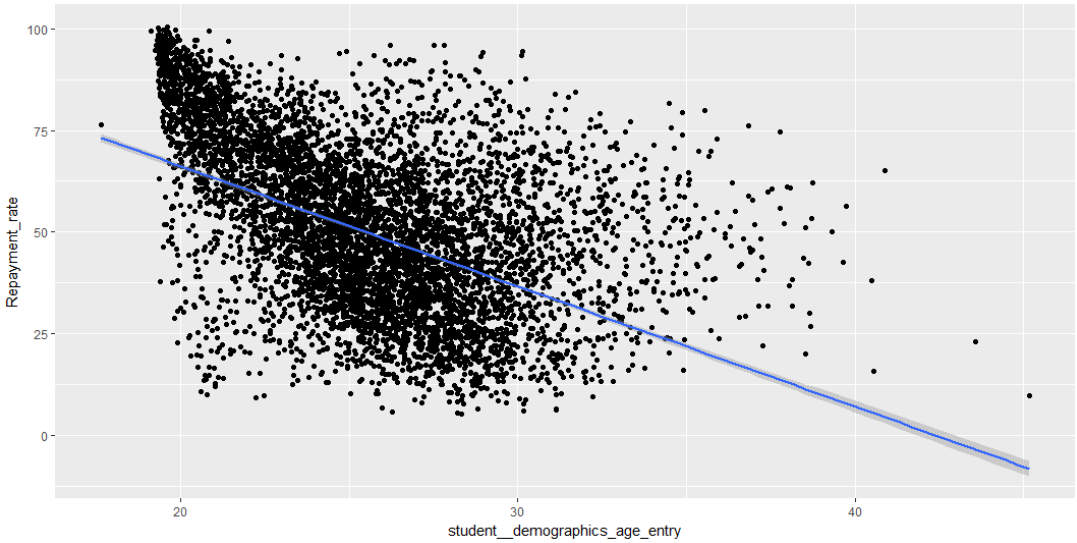

Repayment_rate
100
75
50
25
0
0.25
0.50
0.75
student__share_firstgeneration_parents_somecollege

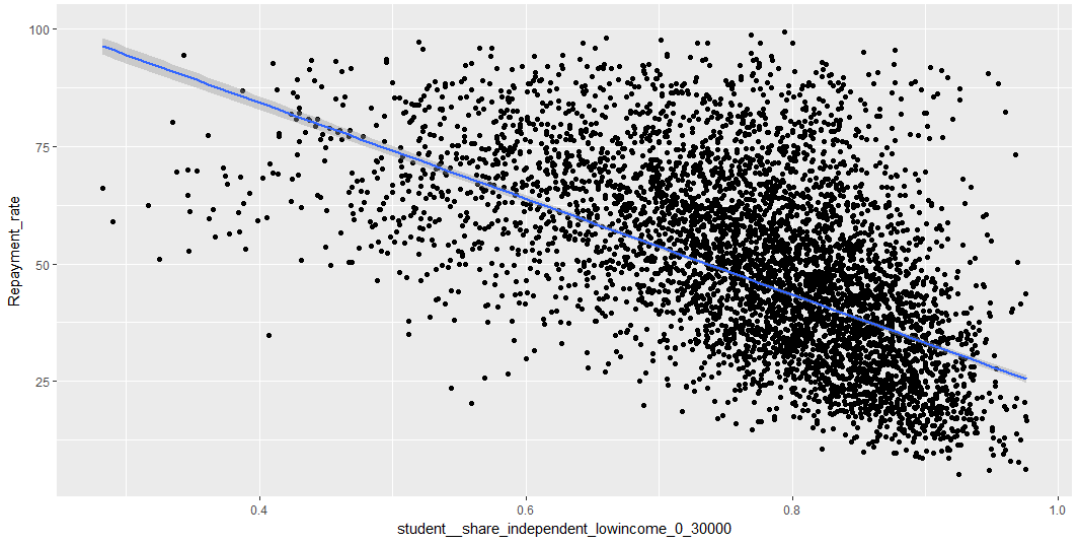

Repayment_rate
90
60
30
0
15
20
25
30
35
admissions__act_scores_75th_percentile_math

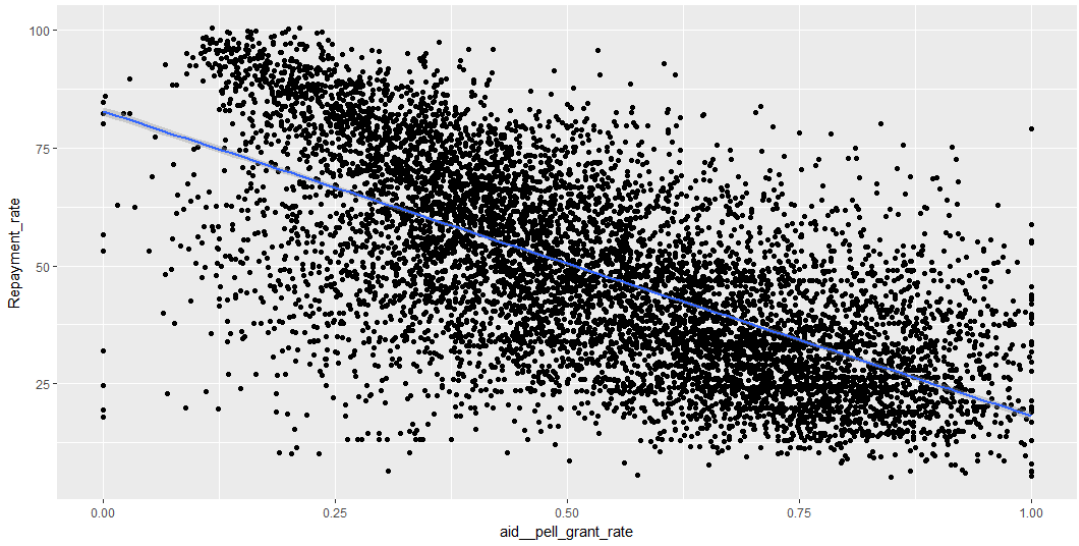

Repayment_rate
90
60
30
0
750
1000
1250
1500
admissions__sat_scores_average_overall

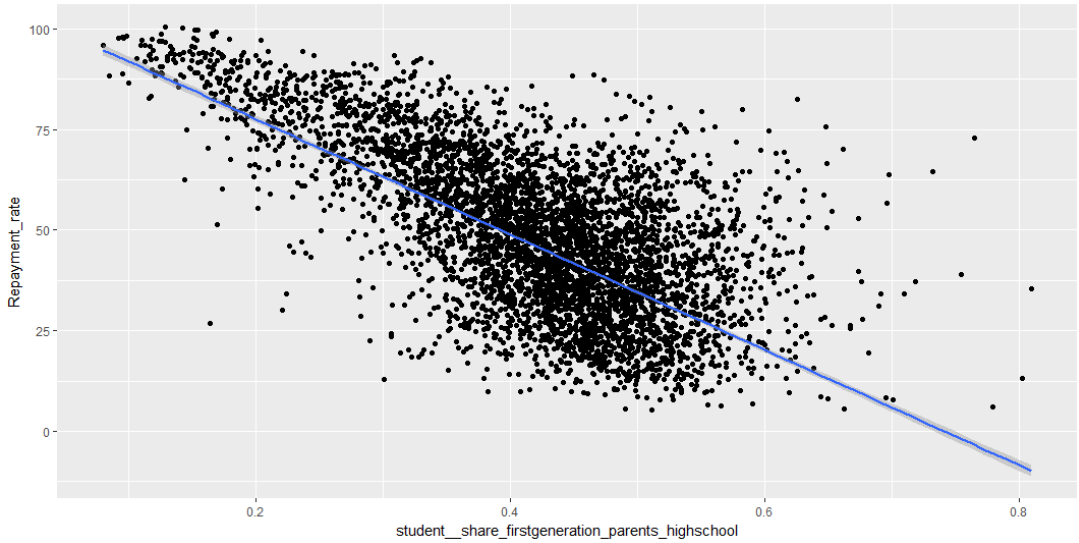

Repayment_rate
100
75
50
25
0.00
0.25
0.50
0.75
1.00
completion__completion_rate_4yr_150nt_pooled

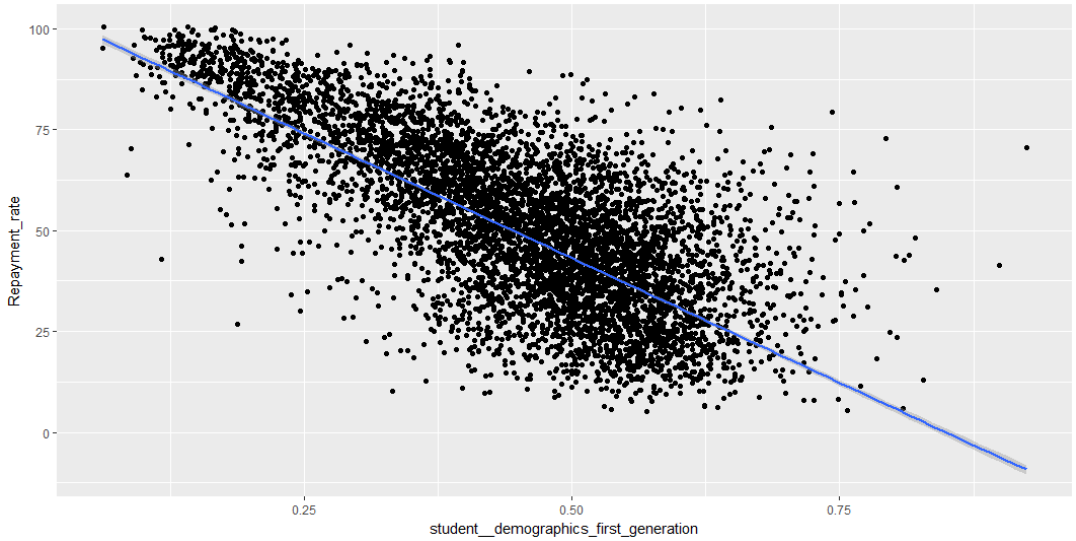

Repayment_rate
90
60
30
0
400
500
600
700
admissions__sat_scores_midpoint_critical_reading

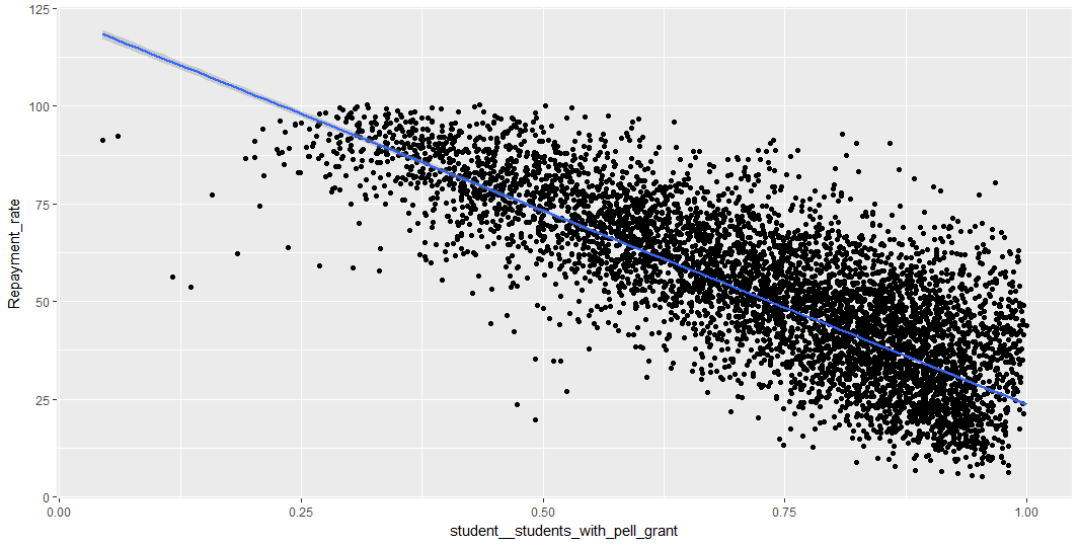

Repayment_rate
100
75
50
25
0
20
30
40
student__demographics_age_entry

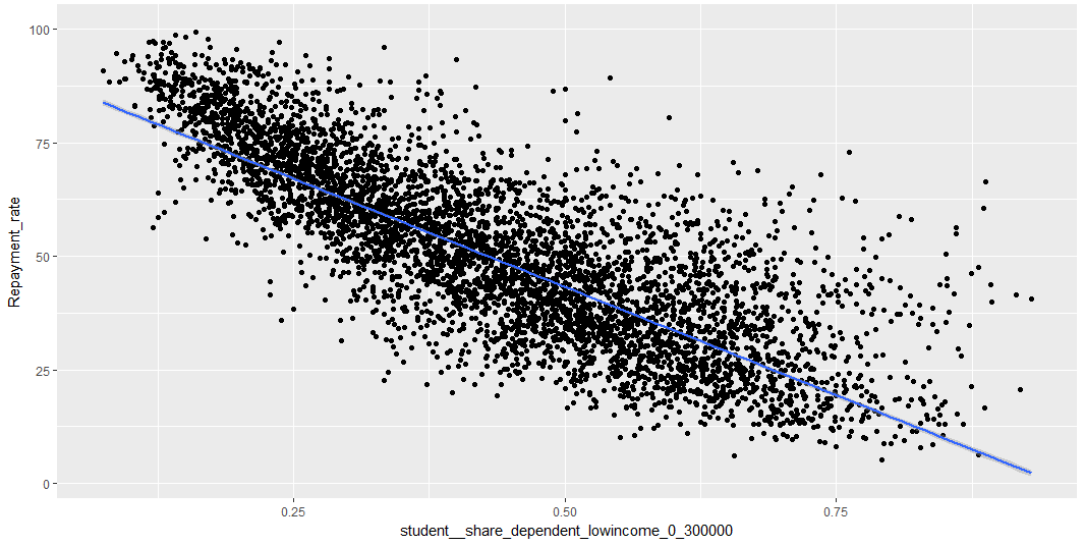

Repayment_rate
100
75
50
25
0.4
0.6
0.8
1.0
student__share_independent_lowincome_0_30000

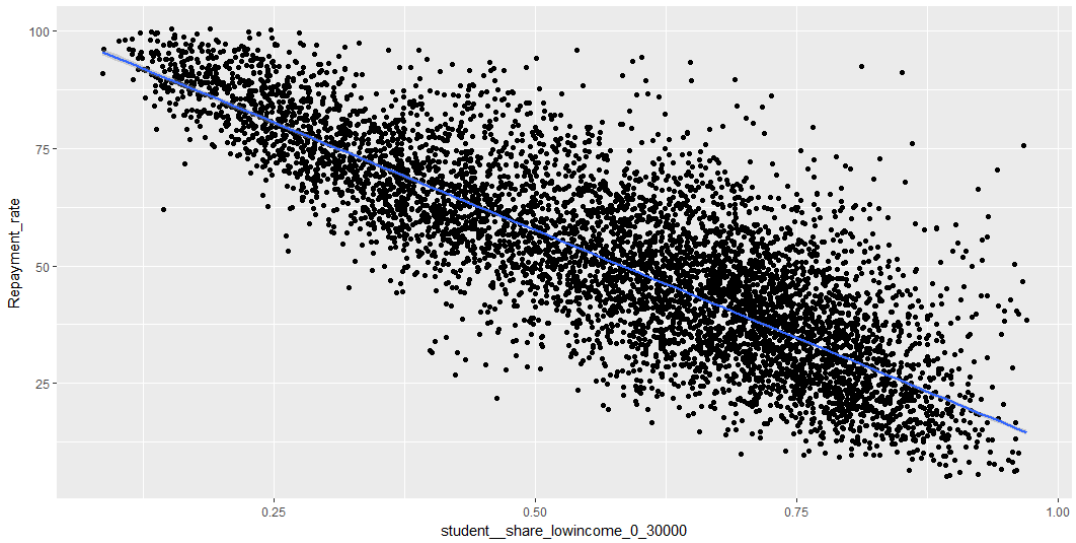

Repayment_rate
100
75
50
25
0.00
0.25
0.50
0.75
1.00
aid__pell_grant_rate


Repayment_rate
100
75
50
25
0
0.2
0.4
0.6
0.8
student__share_firstgeneration_parents_highschool


Repayment_rate
100
75
50
25
0
0.25
0.50
0.75
student__demographics_first_generation


Repayment_rate
125
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00
student__students_with_pell_grant


Repayment_rate
100
75
50
25
0
0.25
0.50
0.75
student__share_dependent_lowincome_0_300000



Based on these plots, some of the points worth noticing are as follows:

- Institutes that admit students with a higher average overall SAT score tend to have a higher repayment rate.
- Institutes that have more aided students with family incomes between $75,001-$110,000 in nominal dollars, tend to have a higher repayment rate.
- Institutes where dependent students have a low median family income or low average family income tend to have a low repayment rate. Similarly, schools with a higher percentage of students who are financially independent and have family incomes between $0-30,000 tend to have a low repayment rate.
- Institutes where the share of students who received a Pell Grant while in school is higher tend to have a low repayment rate.
- Institutes with a higher share of first-generation students tend to have a low repayment rate.
- Institutes with a higher percentage of students whose parents' highest educational level is high school, tend to have a low repayment rate. But Institutes with a higher percentage of students whose parents' highest educational level is some form of postsecondary education tend to have a high repayment rate.
- Institutes with a higher completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion/6 years), tend to have a higher repayment rate.
- Institutes with a higher average age of entry tend to have low repayment rate.

3. Classification of Institutes based on Repayment Rates

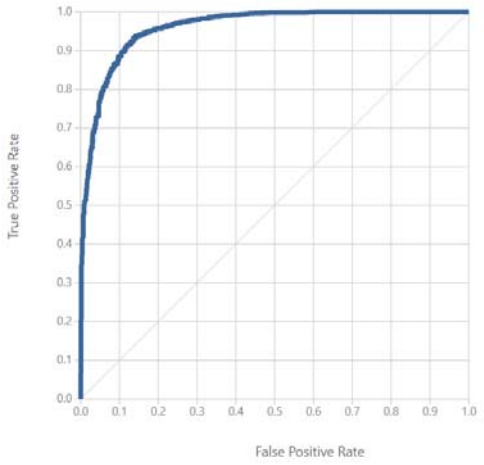
The author also created a predictive model to classify the United States Institutes into two classes:

- Below- Average (Repayment rate less than average repayment rate, i.e. 47.371)
- Above-Average.

The model was created using the Two-Class Logistic Regression model. The model was trained with 70% of the data, and tested with the remaining 30%. The following results were obtained:

True Positive	False Positive	False Negative	True Negative
1213	155	99	997

The Received Operator Characteristic (ROC) curve for the model is shown below, where the blue curve indicates the model's performance, and the diagonal line shows the outcome of a random guess. The Area-under-the-Curve turns out to be 0.960.

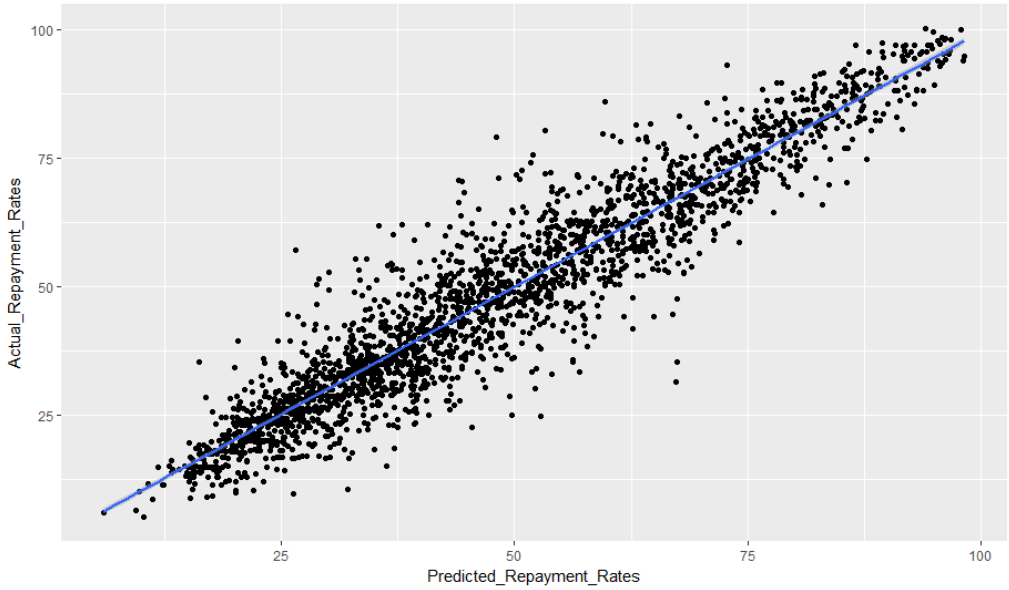


Some of the important performance measures for this classification model are provided below:

Accuracy	Recall	Precision	F1 Score
0.897	0.925	0.887	0.905

4. Regression model for Prediction of Repayment Rates

Finally a regression model was created to predict the repayment rates. Several models were created and it turned out that Boosted Decision Tree Regression model gave best results in our case. The model was trained with 70% of the data, and tested with the remaining 30%. A scatter plot showing the actual and predicted repayment rates is shown below:



The above plot shows a clear linear relationship between the predicted and actual repayment rates in the dataset. The Root Mean Squared Error (RMSE), the performance metric considered in this regression problem, turns out to be 6.812851.

5. Conclusion

The analysis carried out in this work shows clearly that we can confidently predict the students loan repayment rates from the characteristics like the school they went to, the degree they earned, the area they live in, their family circumstances, and how much money they borrowed.