Analysis of student loan repayment rate

Executive Summary

The goal of this analysis is to predict the percent of students that are actively repaying their student loans within three years of graduating. This is referred to as the repayment rate for student loans in U.S. Students use their education to get jobs where they can afford to pay back the loans that they took out. The predictions are based on information about educational institutions, degrees schools offers, financial makeup of the student population, SAT scores, academic merits of the school, graduation rates, and additional demographic information. This analysis is based on 8705 observations.

Initial step involved exploring data, gathering summary statistics and creating visualizations of data to identify potential relationships between repayment rate and features relating to type of school, school academic programs, aid information, completion rate, cost of tuition, degrees awarded, school region, student demographics, sat scores, students' family income and more. After exploring the data, a regression model was created to predict the repayment rate for student loans.

After performing the analysis, in conclusion, while many factors can help predict repayment rate, significant features found in this analysis were:

- **Student median family income**: Students with higher median family income tend to have higher repayment rate.
- **Student SAT scores**: Average SAT equivalent scores of admitted students. Students with higher SAT scores tend to have higher repayment rate.
- **Completion rate**: Completion rate for first-time, full-time students at four-year institutions. Students who went to schools with higher completion rate tend to have higher repayment rate.
- **School ownership**: Median repayment rate was higher for private-nonprofit schools than public schools, which in turn was higher than median repayment rate for private-for-profit schools.
- **School region**: Of the 10 regions in U.S., median repayment rate was highest for New England region followed by Mid-East and Plains regions.

Overview of steps

Following steps were performed in this analysis.

1. Data exploration
2. Data preparation and handling missing data
3. Feature selection, scaling, and editing meta data
4. Regression modeling
5. Model evaluation and updates

6. Predicting repayment rate for private test data

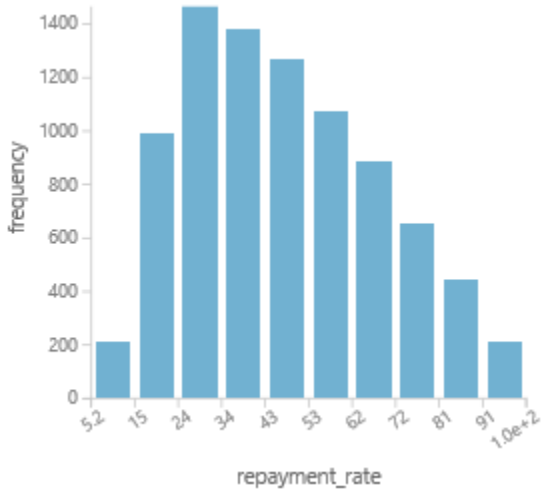
Initial Data Exploration

https://datasciencecapstone.org/competitions/1/student-loans/page/2/ link has description of the data. Appendix A contains description of the features considered in the analysis.

Initial data exploration began with summary and descriptive statistics based on 8705 rows of data. Summary statistics of dependent variable **repayment rate** is as follows.

Mean	47.3709
Median	44.855
Min	5.1627
Max	100.4736
Standard Deviation	20.9876
Unique Values	5594

Repayment rate histogram is as follows. This indicates a right tailed distribution where its mean is greater than the median.



Data contained 443 features/columns relating to type of school, school academic programs, aid information, completion rate, cost of tuition, degrees awarded, school region, student demographics, sat scores, and students' family income. A lot of data was missing. An attempt was made to extract high level features based on broad categories. High level independent variables/features considered were student demographics median family income, sat scores (average overall), school ownership, school region id, completion rate (completion rate for first-time, full-time students at four-year institutions), and admission rate.

Repayment rate versus median family income, sat scores and completion rate

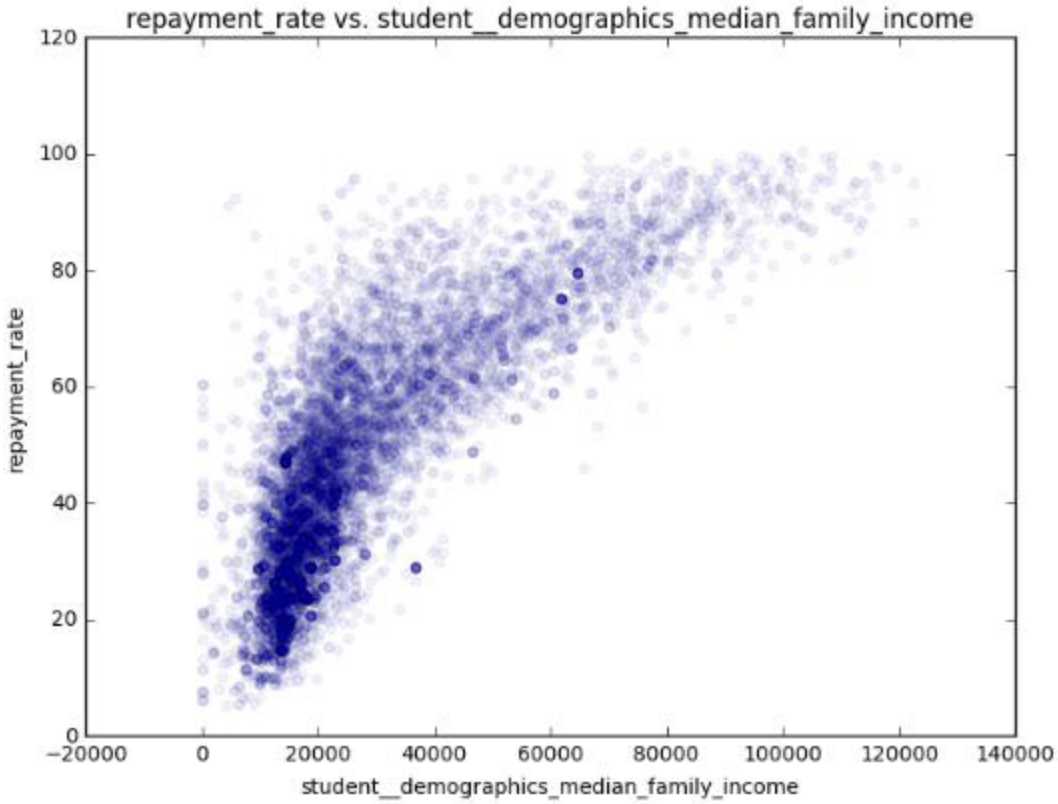
Following table summarizes statistics for median family income and sat scores.

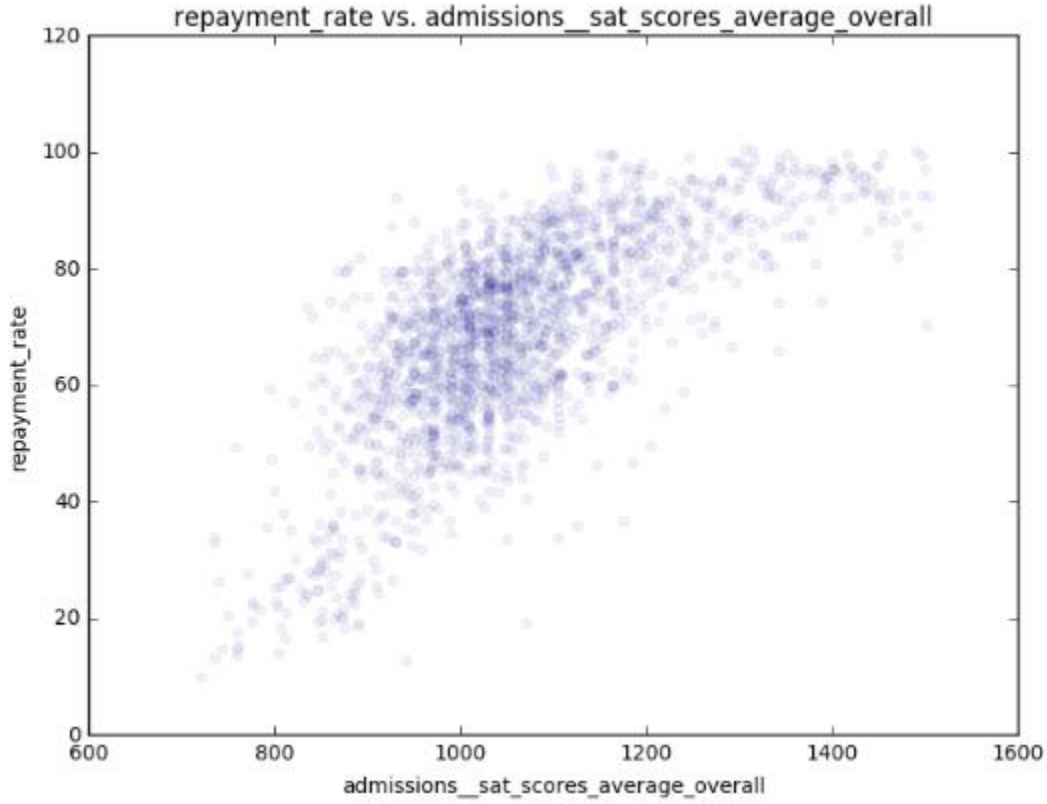
	student__demographics_median_family_income	admissions__sat_scores_average_overall	repayment_rate
count	8697.000000	1889.000000	8705.000000
mean	27979.121912	1058.311805	47.370863
std	19448.175330	130.766517	20.987642
min	0.000000	720.000000	5.162708
25%	15191.782760	973.000000	30.228006
50%	20770.794850	1039.000000	44.855045
75%	34094.508930	1121.000000	62.622899
max	122445.946800	1505.000000	100.473631

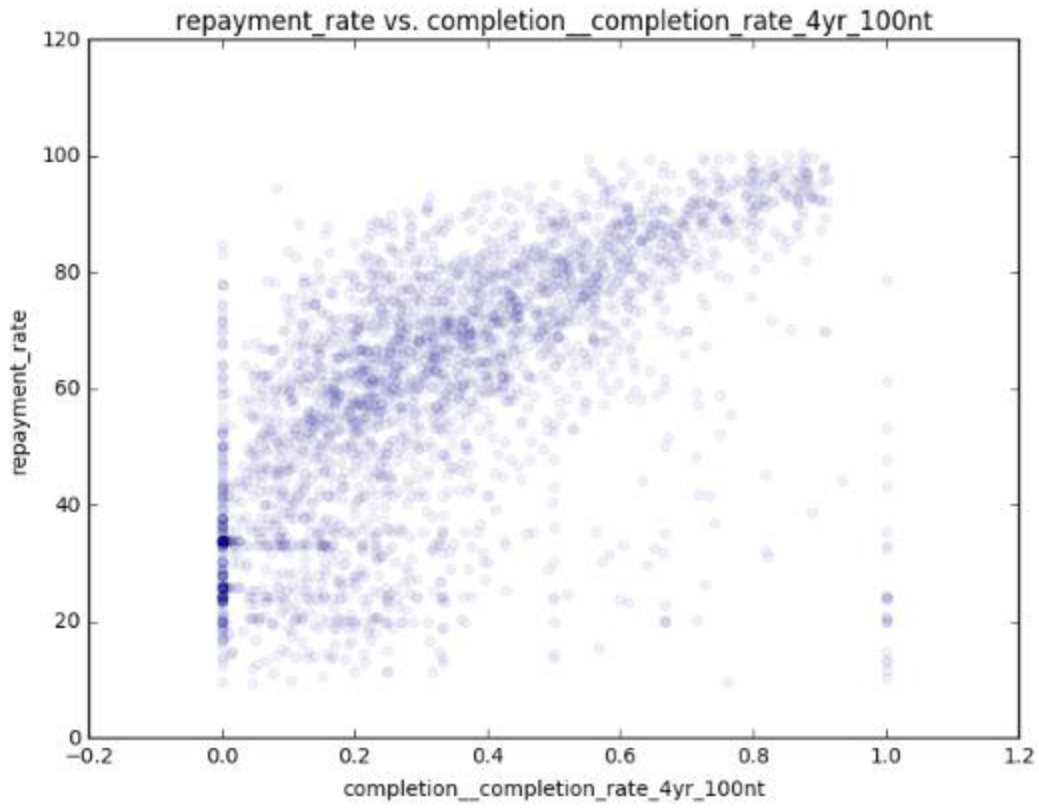
Repayment rate was directly correlated to student demographics median family income, student sat scores (average overall), and completion rate (completion rate for first-time, full-time students at four-year institutions) with the following correlation coefficients.

student__demographics_median_family_income	admissions__sat_scores_average_overall	completion__completion_rate_4yr_100nt
0.800772	0.702053	0.616207

Following scatter plots demonstrate the correlation. Students with higher median family income tend to have higher repayment rate. Students with higher SAT scores tend to have higher repayment rate. Students who went to schools with higher completion rate tend to have higher repayment rate.

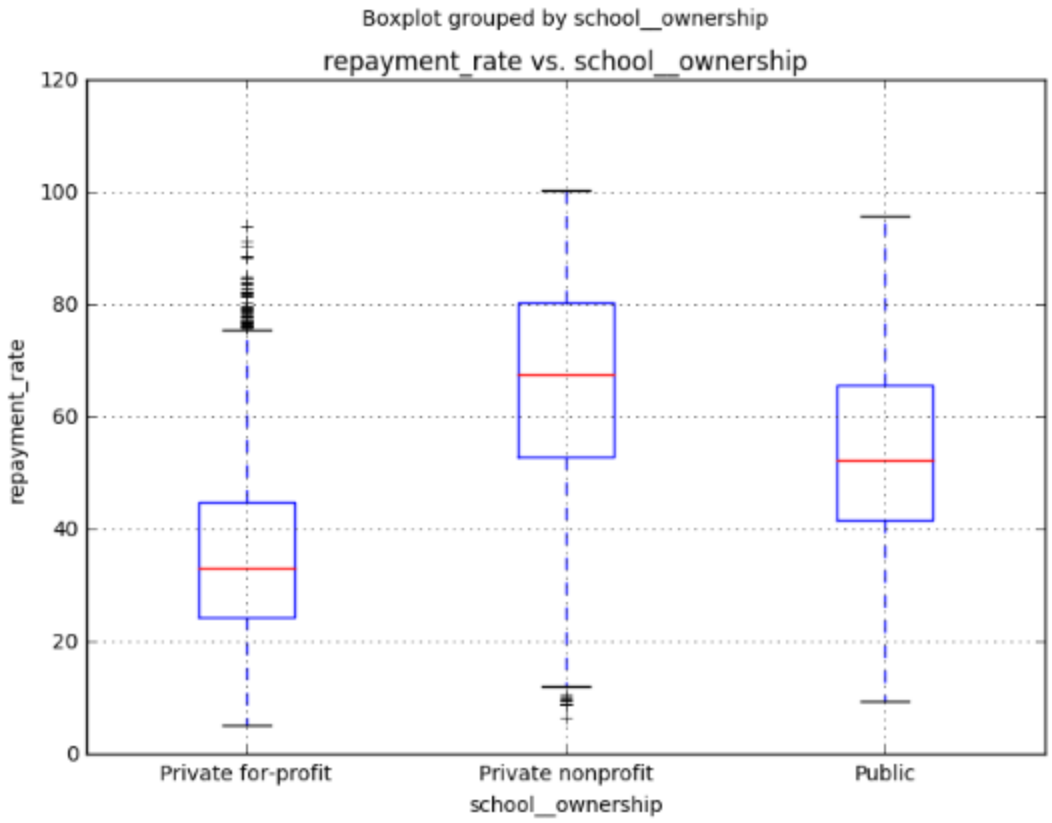

repayment_rate vs. student__demographics_median_family_income
repayment_rate
student__demographics_median_family_income
120
100
80
60
40
20
0
−20000
0
20000
40000
60000
80000
100000
120000
140000


repayment_rate vs. admissions__sat_scores_average_overall
repayment_rate
admissions__sat_scores_average_overall
120
100
80
60
40
20
0
600
800
1000
1200
1400
1600



Repayment rate versus type of school

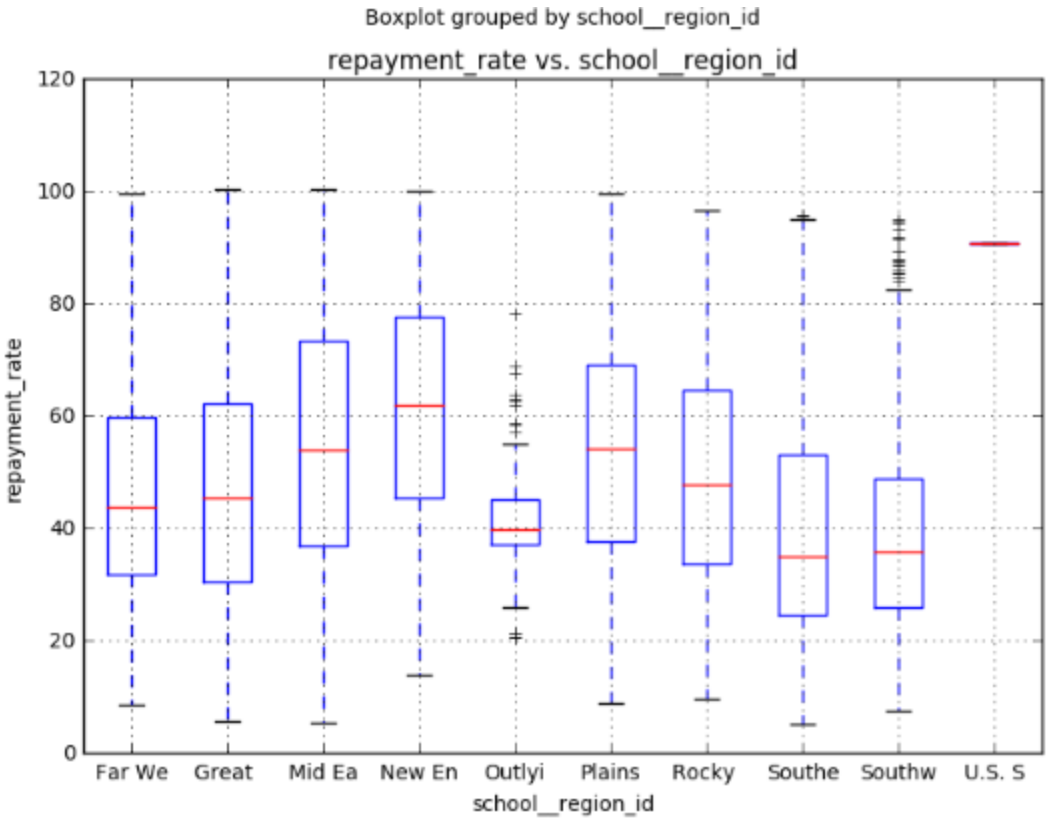
Median repayment rate for private nonprofit schools was higher than public schools which was higher than the median repayment rate for private for-profit schools.



Repayment rate versus school region

Schools were classified in following ten regions. Median repayment rate was highest for New England region followed by Mid-East and Plains regions.

```
['Rocky Mountains (CO, ID, MT, UT, WY)',
       'Plains (IA, KS, MN, MO, NE, ND, SD)',
       'Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV)',
       'Southwest (AZ, NM, OK, TX)', 'Mid East (DE, DC, MD, NJ, NY, PA)',
       'New England (CT, ME, MA, NH, RI, VT)',
       'Far West (AK, CA, HI, NV, OR, WA)',
       'Great Lakes (IL, IN, MI, OH, WI)',
       'Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)',
       'U.S. Service Schools']
```



Repayment rate versus admission rate

Repayment rate had a negative correlation of -0.24 with admission rate implying that schools that were more selective with lower admission rate had higher repayment rate as is inferred from the following plot.

Getting data ready for machine learning

This data has 8705 rows and 443 features/columns. The nature of the columns relating to categorical academic programs and alike makes it sparse with a lot of non-applicable or missing data. As an example, only 1889 rows out of 8705 had average SAT score values. Getting the data ready for machine learning required addressing the non-applicable and missing data. Dropping rows with missing data was not an option as it cut the data down to 20% of the original data. A better approach was to fill all numeric features' missing data with the median for the feature. Then categorical columns with missing data that didn't have correlation with repayment rate were dropped resulting in 431 columns and 8705 rows in the prepared data.

Private unlabeled test data versus labeled data

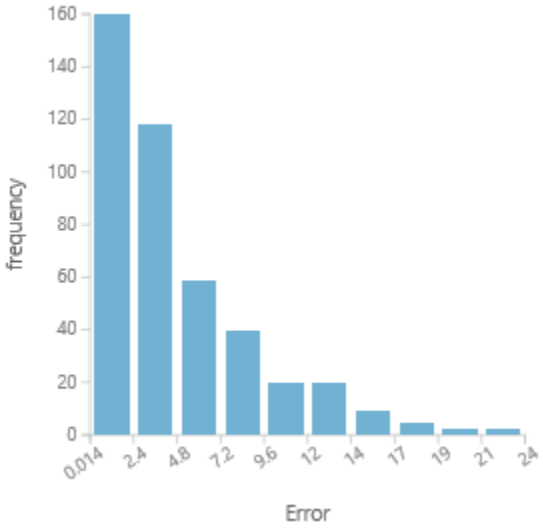
Private unlabeled test data consisted of 6390 rows which is equivalent to 42% of data requiring prediction of repayment rate (label) information. 58% of data had repayment rate (label) information. 58% of the data is equivalent to 8705 rows referred to above. Model training was performed on a subset of the labeled data.

Regression model to predict repayment rate

Prepared data with labels was split into 80% training data and 20% labeled test data for evaluation. Boosted decision tree regression algorithm was used to train the model. The trained model yielded a root mean square error (RMSE) of 6.769 with a coefficient of determination of 0.89 as follows.

Mean Absolute Error	4.882435
Root Mean Squared Error	6.769043
Relative Absolute Error	0.287143
Relative Squared Error	0.109094
Coefficient of Determination	0.890906

Histogram of residual error between the predicted and actual repayment rate values is as follows.



Dimensionality Reduction

Principal Component Analysis (PCA) was used to reduce the feature set from 431 to 50 and 200 features respectively. For both the cases, adding PCA to the pipeline caused the root mean square error (RMSE) metric to go up when evaluated against the 20% labeled test data. Based on this observation, PCA was dropped from the pipeline for the final predictions of the private unlabeled test data.

Predicting repayment rate for private unlabeled test data

The data transformation steps applied to train the model were applied to the private unlabeled test data and passed into the machine learning model to generate scored labels. These scored labels yielded a root mean square error (RMSE) of 7.66 indicating the predictions were within plus/minus 7.66 units/percent of the actual repayment rate.

Conclusion

This analysis has shown that the repayment rate for student loans can be predicted (within plus/minus 7.66 units of actual values) from student and school data, specifically relating to students' median family income, average SAT scores, school completion rate, school ownership (private nonprofit, private for-profit or public), region and other information.

Appendix A

https://datasciencecapstone.org/competitions/1/student-loans/page/2/ has detailed description of data. Following are a subset of features that were considered in the analysis.

- `admissions__admission_rate_overall`: Admission rate
- `aid__federal_loan_rate`: Percent of all federal undergraduate students receiving a federal student loan
- `aid__loan_principal`: The median original amount of the loan principal upon entering repayment
- `admissions__sat_scores_average_overall`: Average SAT equivalent score of students admitted
- `completion__completion_rate_4yr_100nt`: Completion rate for first-time, full-time students at four-year institutions (100% of expected time to completion/6 years)
- `cost__attendance_academic_year: Average cost of attendance (academic year institutions)`
- `student__demographics_median_family_income`: Median family income in real 2015 dollars
- `school__ownership: Control of institution (Possible Values: Private nonprofit, Private for-profit, Public)`
- `school__region_id: Region (IPEDS) (Possible Values: Rocky Mountains (CO, ID, MT, UT, WY), Plains (IA, KS, MN, MO, NE, ND, SD), Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV), Southwest (AZ, NM, OK, TX), Mid East (DE, DC, MD, NJ, NY, PA), New England (CT, ME, MA, NH, RI, VT), Far West (AK, CA, HI, NV, OR, WA), Great Lakes (IL, IN, MI, OH, WI), Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI), nan, U.S. Service Schools)`
- `student__ size`: Enrollment of undergraduate certificate/degree-seeking students
- school__ `instructional_expenditure_per_fte`: Instructional expenditures per full-time equivalent student