Stats 159 Final Project Report

Zhongling Jiang

November 26, 2016

Data Cleaning and Processing

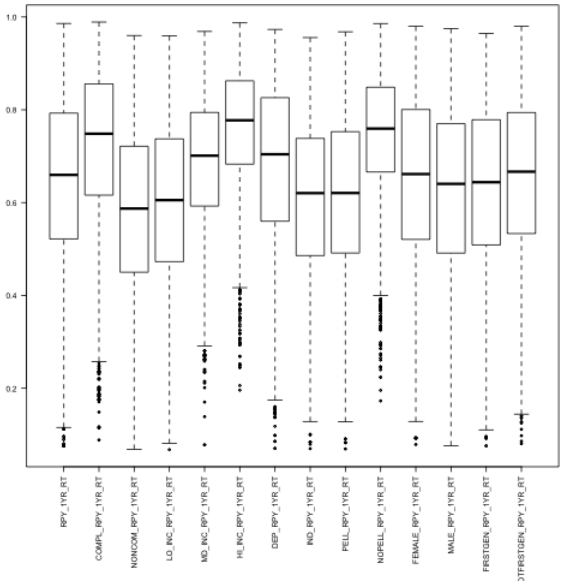
There are three types of missing data in the dataset: NA, NULL and PrivacySuppressed. These values account for measurable percentage of the dataset, so it is not appropriate to simply remove them. A proportion of NULL values are meaningful as they indicate the absence of binary variables, or numerical zero. In order to use these data, we replace NULL values by zero. Meanwhile we replace PrivacySuppressed value by NA.

However, there is still a large percentage of NA present. To solve this problem, we implemented predictive mean matching (pmm) imputing method, which involves training a regression model using other fully observed variables, to predict missing values.

Repayment

One objective of our model is to identify the key indicator of borrowing risk. In repayment section, we observe that there are several important predictors: one- and three-year Cohort Default Rate (CDR), and 1,3,5,7-year Repayment Rate (RPY_YR_RT). There are also data on repayment rate split on different categories e.g. degree completor v.s non-completor, income low v.s medium v.s high, etc.

We first look at Repaymen Rate in different categories of students. Below is the boxplot of one-year repayment rate with regard to categories. It shows the facts that whether students complete degree, students' family income level and whether students receive pell grants will affect the repayment rate significantly. The result is also shown by running two sample t-test on these categories.

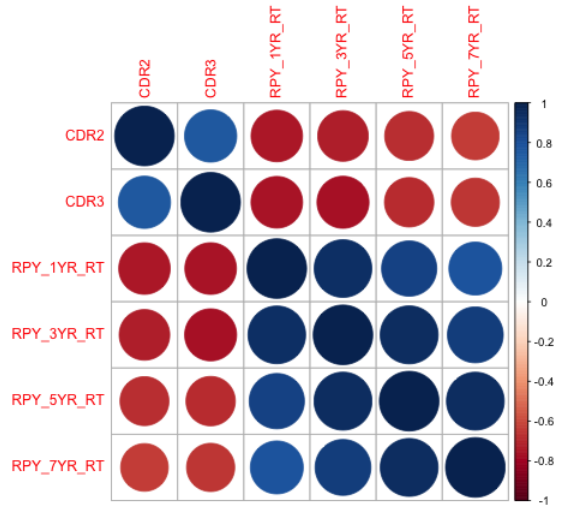


There is also high correlations between default rate and repayment rate. The following is the correlation matrix among 1,3 year CDR and 1,3,5,7 year RPY_RT

```
                CDR2       CDR3 RPY_1YR_RT RPY_3YR_RT RPY_5YR_RT RPY_7YR_RT
CDR2       1.0000000  0.7675589 -0.7467522 -0.7210302 -0.6748901 -0.6380050
CDR3       0.7675589  1.0000000 -0.7501919 -0.7619241 -0.6864179 -0.6541115
```

```
RPY_1YR_RT -0.7467522 -0.7501919  1.0000000  0.9354577  0.8590879  0.7880722
RPY_3YR_RT -0.7210302 -0.7619241  0.9354577  1.0000000  0.9446204  0.8774713
RPY_5YR_RT -0.6748901 -0.6864179  0.8590879  0.9446204  1.0000000  0.9411920
RPY_7YR_RT -0.6380050 -0.6541115  0.7880722  0.8774713  0.9411920  1.0000000
```

And the correlation plot:



Therefore, we choose 3-year CDR as the indicator variable as our model. Firstly, it captures the debt conditon of students in a certain college because students who default on loan is either buried in debt or unable to make enough earnings after graduate. Secondly, it is highly correlated with repayment variables, which measure borrowers ability's pay back the loan. We only need to focus on one of them. We believe CDR3 is an important risk factor that credit instituion needs to consider before they borrow.