Final Project - Predictive Modeling of Credit Risk

Jacqueline Liu, Steven Chen, Zhongling Jiang

December 5, 2016

1 Abstract

College Scorecard dataset is released by U.S Department of Education in September 2015, as a means of providing transparent information for over 7000 degree-granting institutions to students and families who are making decisions concerning recieving secondary education. This information includes the cost of attendance, post-graduate earnings and debt, and financial aid. For credit institutions who are eager to provide more loans to college students, they need to be selective when choosing the 'less risky' colleges to at which they will be giving their loans. In other words, they want to ensure that a college's loan recievers are able to pay loans back after they graduate and will enter the job market rather than default. This paper will explore the correlations between characteristics of colleges and their loan-recieving students' repayment ability. We use predictive modeling with different methods to help credit institutions make informed decisions when evaluating the borrowing risk of each customer.

2 Introduction

We hope to create data models that can make predictions regarding the loan cohort default rate (CDR) based on factors that represents students' earnings and financial aid. In order to better-informed lending decisions, we classify all institutions into high-risk (CDR3 > 0.15)and low-risk (CDR < 0.15) two categories. Our model will tell which category an institution will fall in given some specific metrics. There are two types of model we use to fit the data:

- Logistic Regression
- Gradient Boosting Machine

We will perform the analysis on the data set Most-Recent-Cohort-All-Data-Elements. In each analysis, we will tune the parameters for each respective model using 5-fold cross validation and then evaluate the best model by using the highest classification precision on the test set. A more detailed explanation can be found in *Methods* section.

3 Data

The primary data set we are using is the most recent from College ScoreCard which contains approximately 2000 metrics for 7704 degree-granting institution across U.S. These metrics include demographic data, student academics, costs, student body, financial aid, completion, repayment information, etc. We choose to focus on the most recent year dataset because it contains most up-to-date information; additionally we choose not to use multiple years of data because the features are not independent across years, requiring some sore of normalization or time series analysis outside the scope of our projcet. Among all metrics available, we selected predictors from following three clusters:

- **Repayment**: Loan performance metrics and borrowers' behavior after graduation
- **Financial Aid**: Amount and type of debt that accumulated during schooling
- **Earnings**: Earnings and employment of former students after graduation

Our first task is to clean the dataset, select the right variable to predict, and most relevant predictors by transforming and segementing the dataset. This process is fully detailed in the following *Methods* section.

4 Methods and Analysis

NOTE: our methods drew from a similar study conducted at Stanford University. The paper can be found in the **References** section.

4.1 Preprocessing

Handling missing values

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 all samples associated with 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, and later we will impute these values.

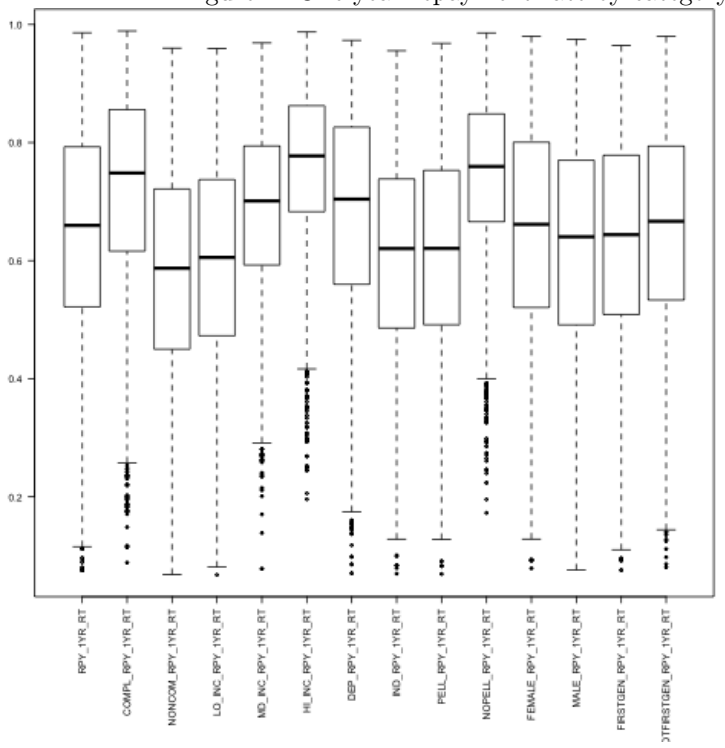
We remove the features that contains over 30 percent of NAs entries, and then remove samples that contains 30 percent of NAs values. That leaves about 5.5k observations. Then we imputed the remaining NA values through the k nearest neighbors (KNN) method using the R package 'VIM'. We were then in possesssion of a full dataset for modeling.

Selecting the response variable

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 Repayment Rate in different categories of students. Figure 1 is the boxplot of one-year repayment rate with regard to categories. It shows 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.

Figure 1: One-year repayment rate by category



There is also high correlations between default rate and repayment rate. Table 1 is the correlation matrix among 1,3 year CDR and 1,3,5,7 year RPY_RT while Figure 2 is a visualization of it.

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 (see links in **References** for more details). Secondly, it is highly correlated with repayment variables, which measure borrowers ability's

	CDR2	CDR3	RPY_1YR_RT	RPY_3YR_RT	RPY_5YR_RT	RPY_7YR_RT
CDR2	1.00	0.77	-0.75	-0.72	-0.67	-0.64
CDR3	0.77	1.00	-0.75	-0.76	-0.69	-0.65
RPY_1YR_RT	-0.75	-0.75	1.00	0.94	0.86	0.79
RPY_3YR_RT	-0.72	-0.76	0.94	1.00	0.94	0.88
RPY_5YR_RT	-0.67	-0.69	0.86	0.94	1.00	0.94
RPY_7YR_RT	-0.64	-0.65	0.79	0.88	0.94	1.00

Table 1: Correlations between Loan Rates

pay back the loan. This high correlations suggests that we only need to focus on one of them. Based on this analysis and our readings, we believe CDR3 is an important risk factor that credit instituion needs to consider before they borrow.

Selecting the predictors

After removing unrelavant features, we narrow our focus down to the Earning section and Financial Aid Section. We observe the correlation plots for each feature group in Figures 3 and 4.

Based on the correlation plots, we retain variables with low correlation and drop those with high correlation. Finally we get a dataset of 11 columns, with the features being: .. and response variable being CDR3.

4.2 Logistic Regression

Logistic regression is one of the most commonly used tools for applied statistics and discrete data analysis. In this model, we have a binary output variable Y, and we want to model the conditional probability

$$Pr\left(Y = 1 \,|\, X = x\right) \tag{1}$$

as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood.

Formally, the model logistic regression model is that

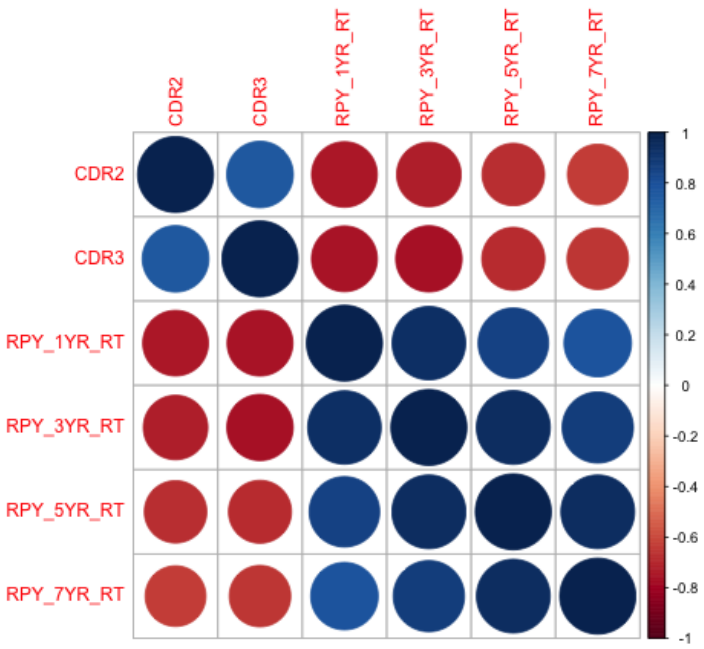
$$log\frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta \tag{2}$$

Solving for p(x), this gives

$$p(x) = \frac{e^{\beta_0 + x\cdot\beta}}{1 + e^{\beta_0 + x\cdot\beta}} \tag{3}$$

To minimize the mis-classification rate, we should predict Y = 1 when p > 0.5 and Y = 0 when p < 0.5. This means guessing 1 whenever $\beta_0 + x \cdot \beta$ is non-negative, and 0 otherwise. Therefore, the decision boundary separating the two predicted classes is the solution of $\beta_0 + x \cdot \beta = 0$.

Figure 2: Correlation Matrix for Loan Rates

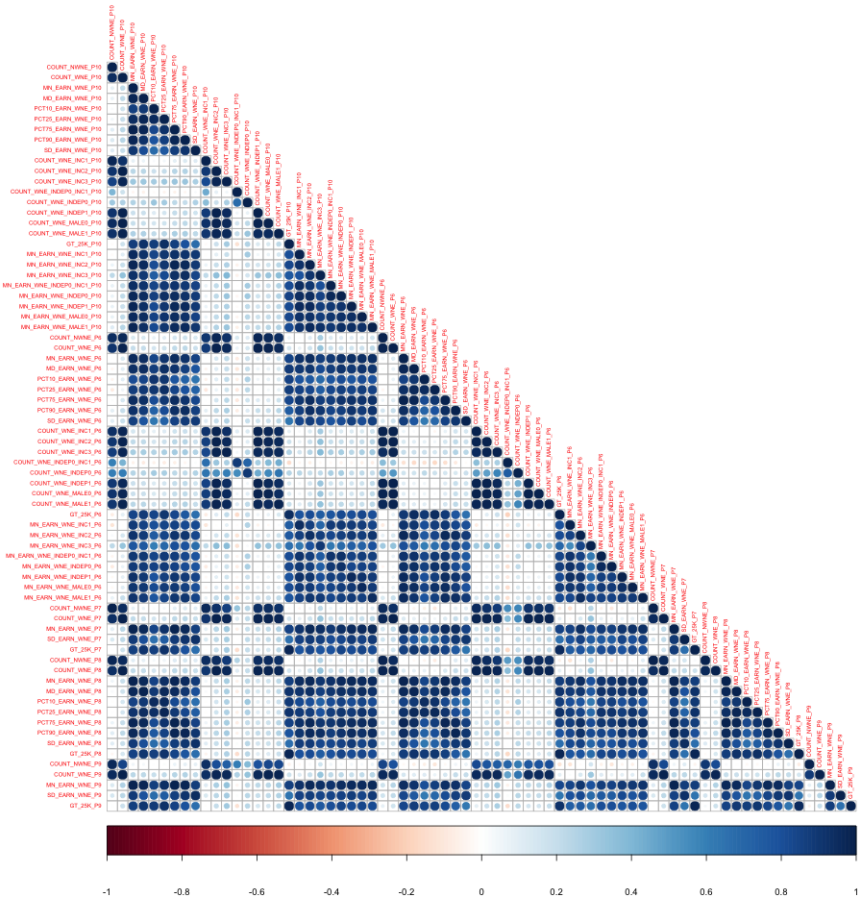


Since Logistic Regression is a classification method, we'll have to transform our label, *CDR3*, into binary labels. Based on the papers we read, we set our threshold at 0.15, which default rates above that considered 'high' risk and those below 'low' risk. We then performed 5-fold cross-validation to gauge the accuracy of logistic regression on this particular problem. The logistic regression itself was achived through the R package 'glm'. We calculated three statistics for each iteration of cross validation: precision, recall, and F1-score (see **Results** for more information on these values).

4.3 Gradient Boosting Machines

We also approached the problem from a regression point of view. The idea behind *gradient boosting* is to combine weak learners in an iterative fashion in order to create a stronger model. Our goal is to find a model M that creates predictions $M(x) = \hat{y}$ for the label (in this case, the three-year default rate) in a way that minimizes the mean-squared error $\sum (y - \hat{y})^2$. We can construct this model by building several smaller models where at each stage i from $i = 0, ..., n$,

Figure 3: Correlation Matrix for Loan Rates



we find the new model by adding an estimator e to the previous model

$$M_{i+1}(x) = M_i(x) + e(x) \approx y$$

$$e(x) = y - M_i(x)$$

This equation then implies that our estimator can be found by modeling the residuals of the previous model.

We used decision trees as our base model, as they are the most popular for this method of learning. This means that ultimately our model is a series of trees that, by modeling the data in this way, is performing gradient descent and is capable of solving multiple regression problems. We used the R package 'gbm' for our gradient-boosted decision trees and performed 5-fold cross validation to

determine our tuning parameters (the number of trees used in our model, the learning rate for each estimator, and the bagging fraction – the percent of our data used to train each tree).

5 Analysis

5.1 Logistic Regression

5.2 Gradient Boosting Machines

<<<<<<< HEAD

6 Results

6.1 Logistic Regression

```
> stats.df = read.csv("../data/logistic-result.csv")
> xtable(stats.df, caption = "Logistic Regression Statistics of 5-Fold Cross Validation")
```

	Precision	Recall	F1
1	0.71	0.81	0.76
2	0.74	0.79	0.76
3	0.74	0.78	0.76
4	0.74	0.78	0.76
5	0.73	0.77	0.75

Table 2: Logistic Regression Statistics of 5-Fold Cross Validation

Precision is the percentage of the positive predictions were correct. Recall is the percentage of positive cases that were correctly classified. Lastly, F1-score considers both precision and recall. The 5-fold cross validation results are displayed in the below table, and on average of the 5-fold, the F1-score is around 76 percent, which is not bad. =======

7 Results

7.1 Logistic Regression

```
> stats.df = read.csv("../data/logistic-result.csv")
> xtable(stats.df, caption = "Logistic Regression Statistics of 5-Fold Cross Validation")
```

Precision is the percentage of the positive predictions were correct. Recall is the percentage of positive cases that were correctly classified. Lastly, F1-score considers both precision and recall. The 5-fold cross validation results are displayed in the below table, and on average of the 5-fold, the F1-score is around 75 percent, which is not bad. >>>>>>> 6f92c175eedb190e39958a0a2b28a9c779b0ccf7

	Precision	Recall	F1
1	0.71	0.81	0.76
2	0.74	0.79	0.76
3	0.74	0.78	0.76
4	0.74	0.78	0.76
5	0.73	0.77	0.75

Table 3: Logistic Regression Statistics of 5-Fold Cross Validation