Uni-X

Regression Analysis on College Scorecard Data

Nov. 29th 2017

Group members: Ahlam Hakami
Bin Luo
Qi Zhang

Goals:

1. What variables can affect repayment rates?

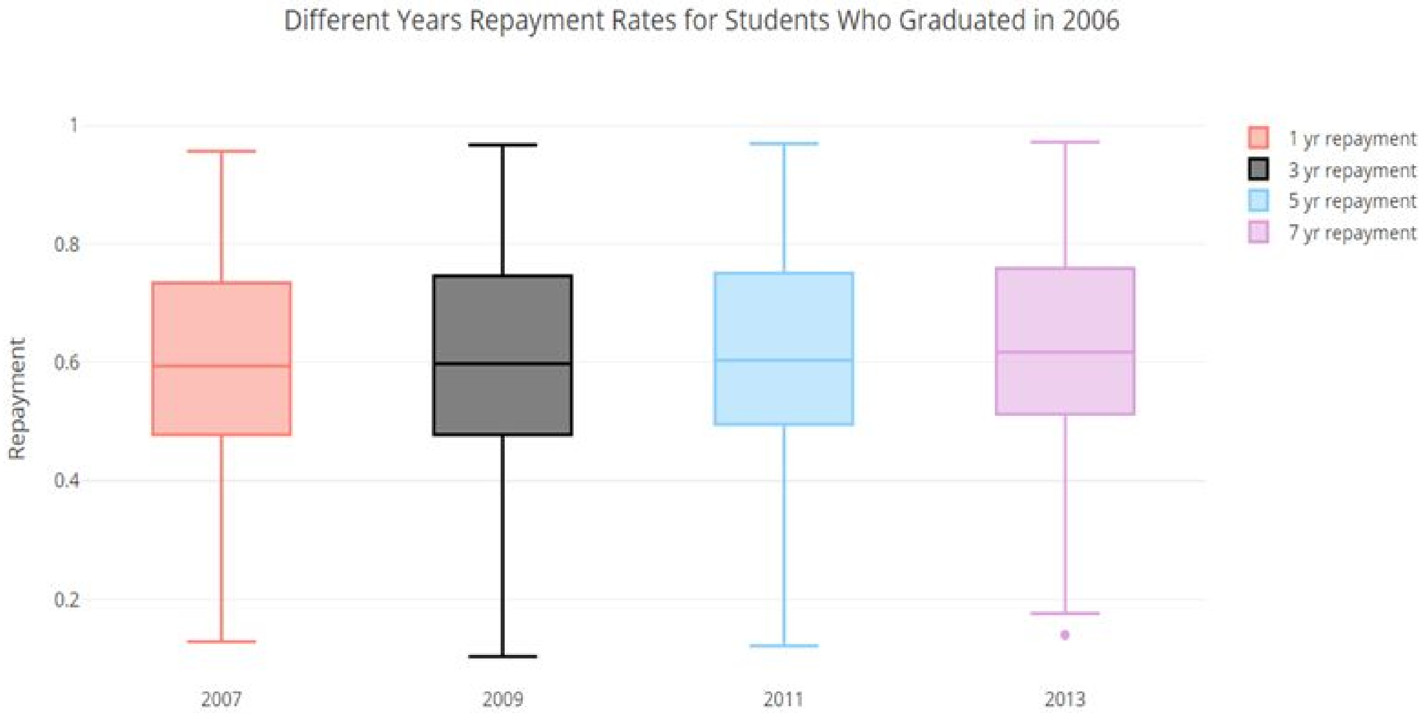
2. Estimate repayment rates based on other variables.

What we have done so far:

1. Performed hypothesis testing on repayment rate.
2. Found highly correlated explanatory variables.
3. c by correlation and standard deviation.
4. Applied PCA on dimension reduction for each category.
5. Fitted a regression model using selective principal components.

Compare male and female repayment rate

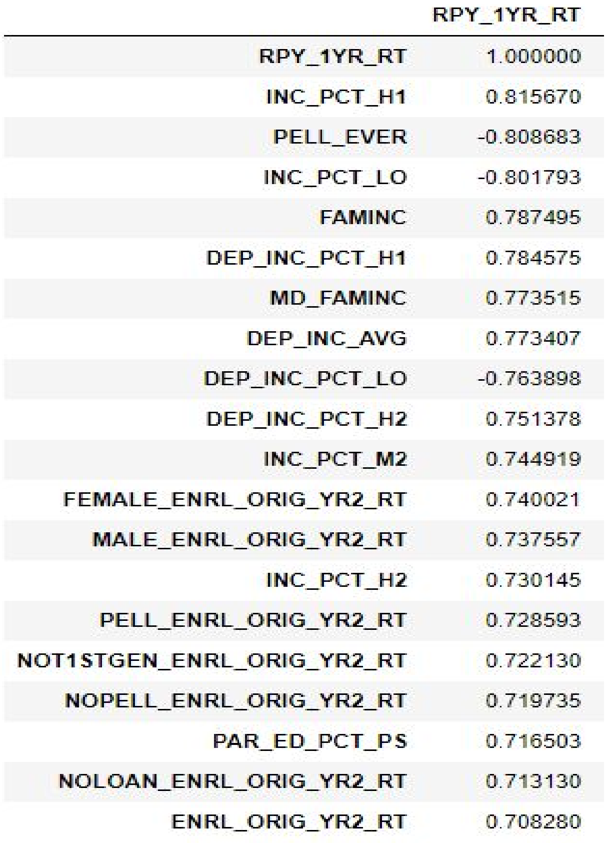
	T value	P value
1 Year	-2.4751	0.0133
3 Years	-1.5119	0.1306
5 Years	-0.3821	0.7024
7 Years	0.0588	0.9531



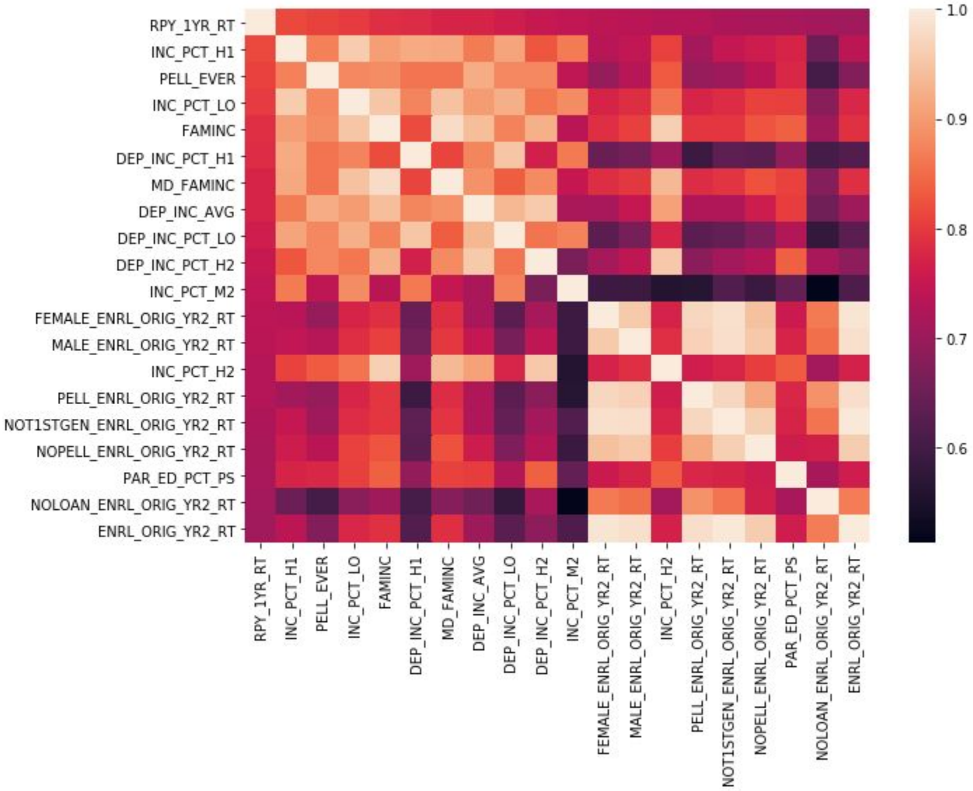
Different Years Repayment Rates for Students Who Graduated in 2006

Since the different years repayment rates are about the same for the same group of people, we only focus on 1 year repayment rate.

Variable Screening

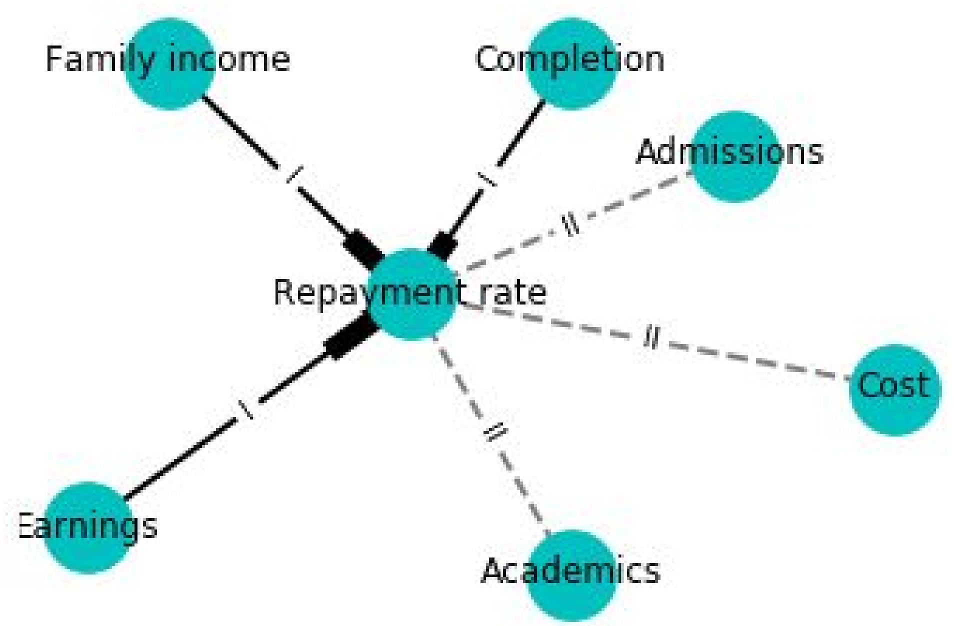


	RPY_1YR_RT
RPY_1YR_RT	1.000000
INC_PCT_H1	0.815670
PELL_EVER	-0.808683
INC_PCT_LO	-0.801793
FAMINC	0.787495
DEP_INC_PCT_H1	0.784575
MD_FAMINC	0.773515
DEP_INC_AVG	0.773407
DEP_INC_PCT_LO	-0.763898
DEP_INC_PCT_H2	0.751378
INC_PCT_M2	0.744919
FEMALE_ENRL_ORIG_YR2_RT	0.740021
MALE_ENRL_ORIG_YR2_RT	0.737557
INC_PCT_H2	0.730145
PELL_ENRL_ORIG_YR2_RT	0.728593
NOT1STGEN_ENRL_ORIG_YR2_RT	0.722130
NOPELL_ENRL_ORIG_YR2_RT	0.719735
PAR_ED_PCT_PS	0.716503
NOLOAN_ENRL_ORIG_YR2_RT	0.713130
ENRL_ORIG_YR2_RT	0.708280



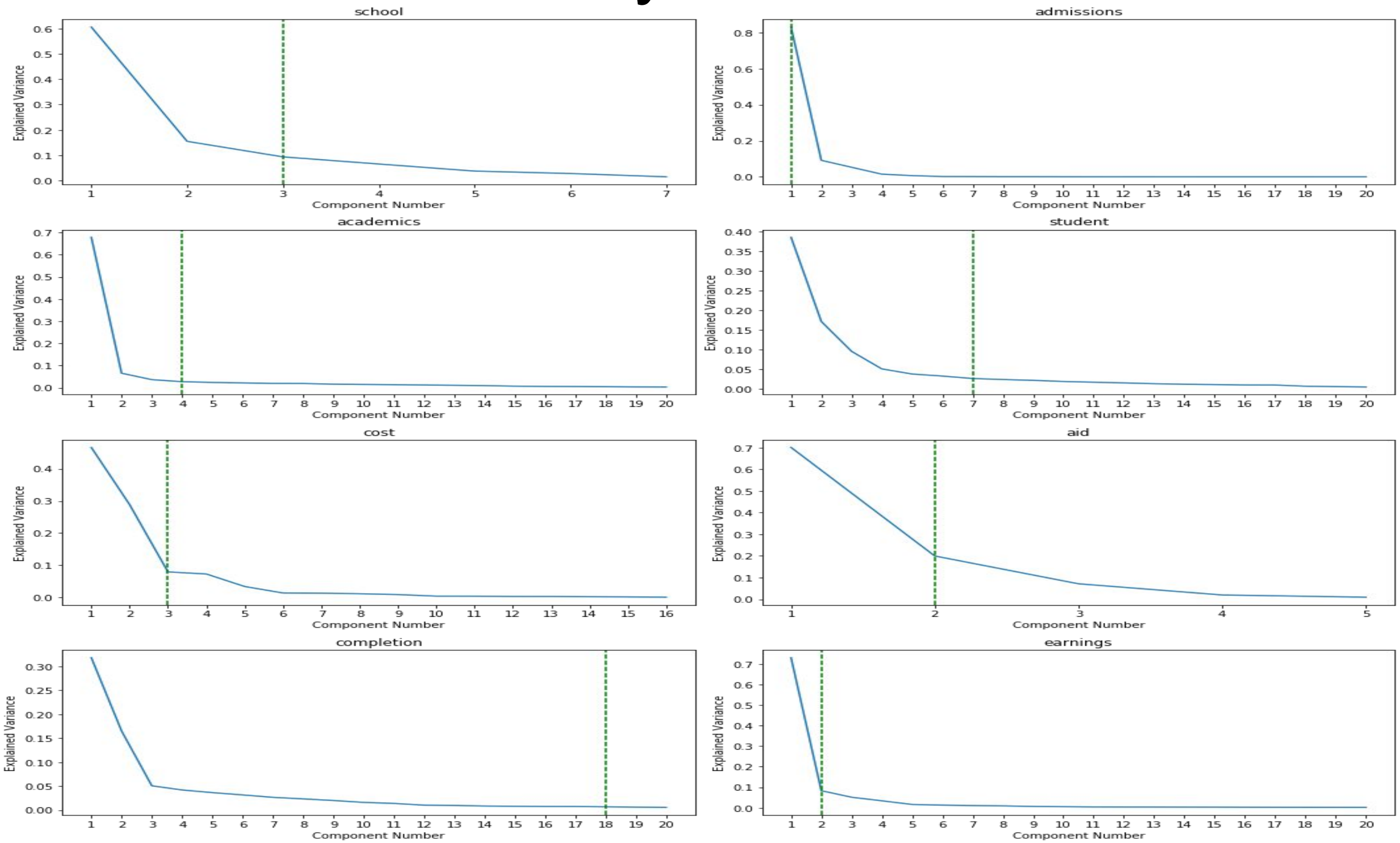
Deleted variables with small standard deviation (< 0.15)

Variable Name		**Variable Name**	
PELL_EVER	Share of students who received a federal loan which in school	FEMALE_ENRL_ORIG_YR2_RT	% of female students who were still enrolled at original institution within 2 years
INC_PCT_LO	% of aided students whose family income is $0-30,000	MALE_ENRL_ORIG_YR2_RT	% of male students who were still enrolled at original institution within 2 years
FAMINC	Average family income in real 2015 dollars	PELL_ENRL_ORIG_YR2_RT	% of students who received a Pell Grant and who were still enrolled at original institution within 2 years
MD_FAMINC	Median family income in real 2015 dollars	NOT1STGEN_ENRL_ORIG_YR2_RT	% of not-first-generation students who completed within 2 years at original institution
DEP_INC_AVG	Average family income of dependent students in real 2015 dollars	NOPELL_ENRL_ORIG_YR2_RT	%of students who never received a Pell Grant at the institution and who were still enrolled at original institution within 2 years
DEP_INC_PCT_LO	% of students who are financially dependent and have family income $0-30,000	NOLOAN_ENRL_ORIG_YR2_RT	% of students who never received a federal loan at the institution and who were still enrolled at original institution within 2 years



We have very similar variables for each feature, a PCA was conducted to deduce the dimension.

Dimension Reduction by PCA



Linear Regression

Response variable: $\log(y/(1-y))$, where y is repayment rate

Explanatory variables: 40 components from all categories

Total number of observation: 35027

R-squared=0.829

Cross-validation MSE: 0.766 (on transformed response variable)

Model Evaluation

Try to plot the validation curve with different variance threshold on PCA...

Questions?