Evaluation of Cost of Postsecondary Education and Post-Graduation Earnings

Paul Previde
CSC869 Spring 2017

Introduction to Problem

- Every year, tens of millions of Americans decide on postsecondary education [1]
- Obama administration published the College Scorecard web site to help prospective students answer this question [2]
- Does attending a cost-competitive institution lead to reduced earning power later?
- What university-level factors affect earning power?

College Scorecard Dataset

Department of Education collects information about universities and students who participate in federal aid programs:

- student demographic, financial and academic performance information
- tuition, admission rates, class sizes

Department of Treasury linked that student information to their tax records

The College Scorecard dataset aggregates and anonymizes student information for each university in each year

College Scorecard Dataset

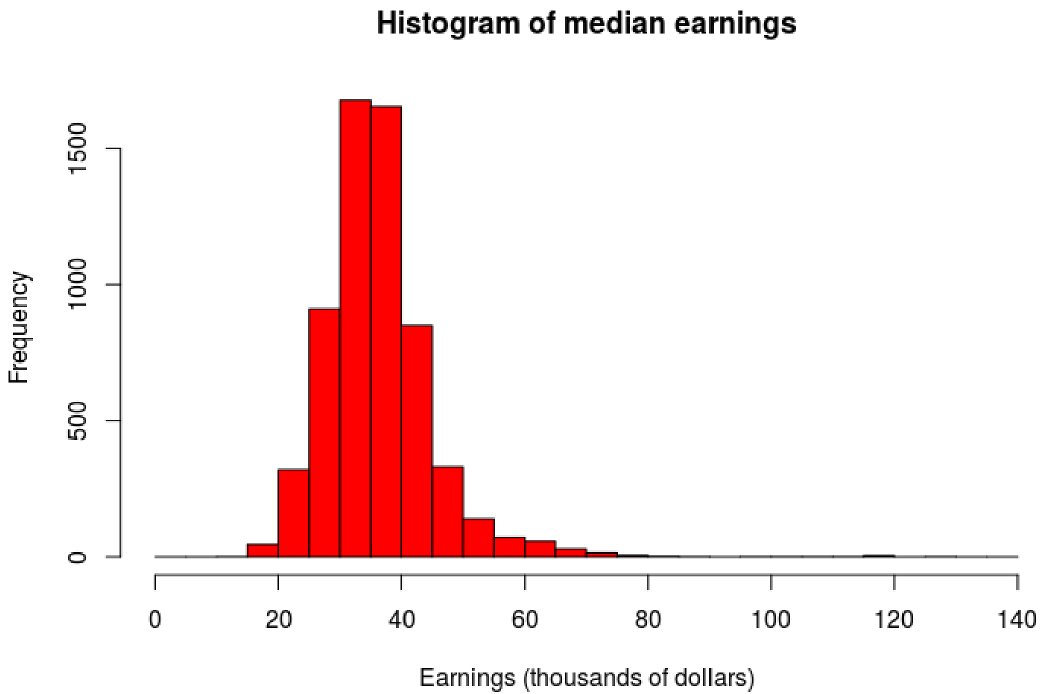
- 124,699 rows by 1,731 columns
- Covers 1996-2013
- 20,601 distinct institutions are represented
- Lots of missing values: 739 ± 421 out of 1,731 variables per observation (~43% of the cells)

Strategies

Use median earnings as the dependent variable

Compare discretization approaches:

- Leave median earnings as a continuous variable
- Discretize into two classes
- Discretize into four classes



Strategies (continued)

Compare missing value handling approaches:

- Omit the observation
- Replace missing values with attribute mean
- Replace missing values with attribute mean for that class

Choose independent variables with fewer missing values where possible

Evaluation Strategy

Compare and interpret results of classification and regression techniques:

Regression
Linear
Lasso
Ridge
Elastic net
Polynomial (second-order)
Polynomial (third-order)

Classification
Naive Bayes
Linear discriminant analysis
Quadratic discriminant analysis
Bagging
Random forest
Adaboost

10-fold cross validation used for all models

All techniques performed using R

Lasso and Model Shrinkage

Lasso can drive model coefficients to zero:

```
(Intercept)    1.033911e+00
admit         -1.178067e+00
satmath        5.783661e-02
satverbal      .
tuitionin      .
tuitionout     1.183743e-01
pricecombined  7.961140e-02
studentbody    3.808343e-09
cs             1.227716e+01
bio           -3.704095e+00
math           .
business       4.417756e+00
lit           -3.417594e+01
```

Test MSE:
Linear regression: 48.56
Lasso regression: 48.94

Regression Results

- Polynomial regression yielded better test MSE than linear models
- Imputation of missing values led to worse MSE

Regression Technique	Test MSE (missing values not imputed)	Test MSE (missing values imputed)
linear	48.56022	94.08526
ridge	48.93166	94.09596
elastic net	48.94044	94.09955
lasso	48.94994	94.10378
polynomial (degree 2)	45.66559	**92.84564**
polynomial (degree 3)	**45.11173**	361.93108

Imputation of Missing Values

Imputation of missing values can lead to a worse-fitting model:

Regression Technique	Adjusted R^2 (missing values not imputed)	Adjusted R^2 (missing values imputed)
linear	0.401	0.211
polynomial (degree 2)	0.450	0.241
polynomial (degree 3)	0.470	0.253

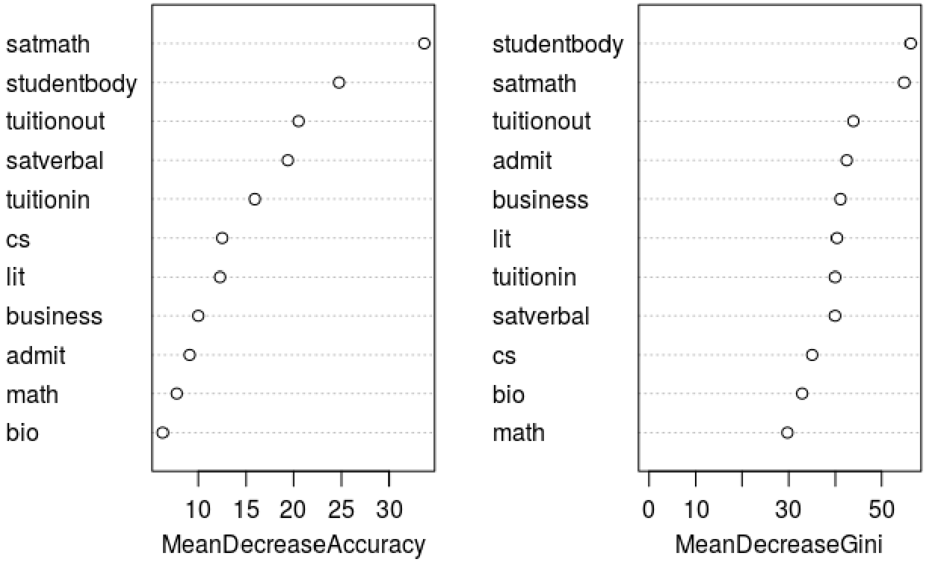
Classification Results

- Tree-based methods offered lower test error rates
- 2-class discretization offered lower test error rates than 4-class
- Imputing missing values led to far lower test error rates

Classification Technique	Error rate (2 classes)		Error rate (4 classes)	
	no imputation of missing values	missing values imputed	no imputation of missing values	missing values imputed
naive Bayes	0.3354037	0.06937275	0.5718862	0.49038965
LDA	0.2969669	0.05566731	0.5496368	0.27540958
QDA	0.3266437	0.06251522	0.5614421	0.35228067
bagging	0.2693425	0.02197383	0.5023611	0.07922838
random forest	0.2720668	0.02243743	**0.4966764**	**0.07891408**
Adaboost	**0.2615661**	**0.02050847**	0.5357201	0.08687812

Variable Importance

Random forest variable importance plots:



Summary of Results

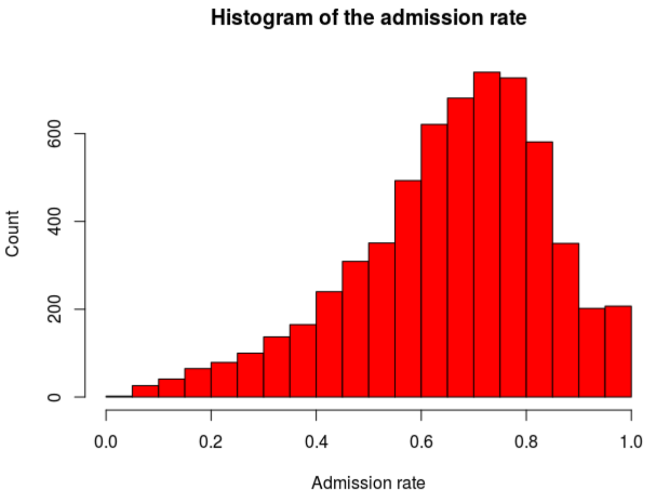
- No significant evidence was observed that spending less for college leads to reduced earning power

- Other university-level factors for which more evidence was found of a relationship with earning power:
  1. university admission rate
  2. average SAT math score
  3. university size

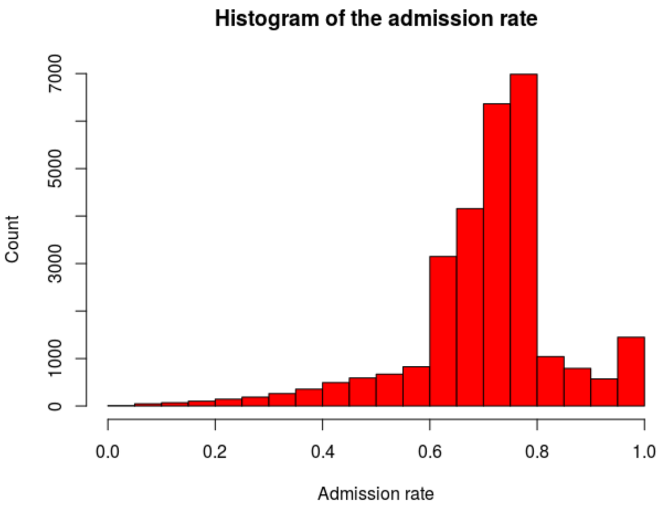
Limitations of the Strategies

Imputation of missing values led to:

- Lower adjusted R^2
- Higher test MSE
- Severe distortion of the variable:



→



Limitations of the Strategies

- Discretization into more classes led to worse classification performance
- Not all the regression and classification models are interpretable
- At best, each result represents further evidence that you have to evaluate

Future Directions

- Employ formal feature subset selection to identify best variables and improve model fit
- Consider alternative approaches to missing value imputation
- Explore other sources of data to complement College Scorecard and alleviate missing value prevalence

The KDD Process

Background research ……….. 10%

Get to know your data ………. 10%

Data preprocessing …………. 50%
(cleaning, discretization,
feature selection)

Data analysis ………………… 20%

Postprocessing ……………… 10%

References

1. “Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education,” U.S. Department of Education, January 2017 (published at: <https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf>, accessed March 10, 2017.
2. College Scorecard web site (<https://collegescorecard.ed.gov>), accessed March 10, 2017.
3. Kaggle web site (<https://www.kaggle.com/college-scorecard>), accessed March-May, 2017.
4. “Better Information for Better College Choice & Institutional Performance,” U.S. Department of Education, January 2017 (published at: <https://collegescorecard.ed.gov/assets/BetterInformationForBetterCollegeChoiceAndInstitutionalPerformance.pdf>, accessed March 10, 2017.