Feature Selection of Post-graduation Income of College Students in the United States

Ewan Wright[1(✉)], Qiang Hao[2], Khaled Rasheed[3], and Yan Liu[4]

[1] University of Hong Kong, Pokfulam, Hong Kong SAR
etwright@hku.hk
[2] Western Washington University, Bellingham, WA, USA
[3] University of Georgia, Athens, GA, USA
[4] University of British Columbia, Vancouver, Canada

Abstract. This study investigated the most important attributes of the 6-year post-graduation income of college graduates who used financial aid during their time at college in the United States. The latest data released by the United States Department of Education was used. Specifically, 1,429 cohorts of graduates from three years (2001, 2003, and 2005) were included in the data analysis. Three attribute selection methods, including filter methods, forward selection, and Genetic Algorithm, were applied to the attribute selection from 30 relevant attributes. We discuss how higher numbers of students in a cohort who grew up in Zip code areas where over 25% of the population hold a Professional Degree was predictive of more college graduates being classified as High income.

Keywords: Attribute selection · Feature selection
Post-graduation income classification · Post-graduation income prediction
Social stratification

1 Introduction

Higher education is an excellent "investment" that should be encouraged by families, schools, communities, and policy makers. The returns of a college degree vis-à-vis a high school diploma has expanded considerably in recent decades. Autor [1] found that this "graduate premium" doubled in real terms between 1979 and 2012. The gap in earnings between the median college educated worker and the median high-school educated worker increased from $17,411 to $34,969 for men, while also increasing from $12,887 to $23,280 for women. Research by Chetty et al. [2] underscores the role of higher education as a key pathway to intergenerational social mobility in the U.S. Further, Hout [3] contends that higher education "makes life better" through a host of social benefits in community relations, health, family stability, and social connections.

Yet as higher education participation has expanded [4], college graduates have become an increasing heterogeneous population with increasingly disparate labor market outcomes [5, 6]. While some graduates are highly successful, others face challenges to gainful employment. Data from the Federal Reserve Bank of New York [7] shows that 43.4% of college graduates aged between 22 and 27 graduates are under-employed or employed in a job that "typically does not require a college degree",

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 38–45, 2018.
https://doi.org/10.1007/978-3-319-93372-6_4

while 12.7 are employed in "low-wage jobs" that tend to pay below $25,000 per annum. Research has established that field of study [8] and institutional selectivity [9] are important features in post-graduation incomes. Building on the literature, this study explored the most important attributes of 6-year post-graduation income of college graduates who used student aid from the U.S. Department of Education, and to what extent of accuracy the select attributes can be used to classify post-graduation income. The research questions were: (1) What are the most important attributes of post-graduation income of college students who graduate with debt repayment obligations? and (2) To what extent can the selected attributes classify post-graduation income of college students who graduate with debt repayment obligations?

2 Research Design

2.1 Data Collection

The data for this study was the latest dataset – released in October 2015 – by College Scorecard under the U.S. Department of Education [10]. This dataset only covered students who used financial aid during their college study period. Each row in the data stands for a student cohort admitted to a certain university. The data ranged from 1996 to 2013, but the 6-year post-graduation income data are only available for the years 1997, 1999, 2001, 2003 and 2005. The response variable in the present study is the mean value of the 6-year post-graduation income of a student cohort. Attributes were filtered based on domain knowledge. Factors deemed less relevant were excluded, such as latitude of the institution and percent of students who passed away within 6 years after graduation.

30 potential attributes (see Appendix A) under five groups were included in this study. The groups are: (1) School, (2) Admission, (3) Cost, (4) Student Cohort, and (5) Socioeconomic Status of Students. Some attributes in certain groups are not available before 2000, such as admission rate in the Admission Group. Thus, only three years of data, including 2001, 2003, and 2005 were used. 1,429 cohorts were included for the data analysis. The response variable, mean income value of each cohort, was discretized into four classes based on the American Individual Income Distribution; including Very low (0 to 25,000), Low (25,000 to 37,500), Middle (37,500 to 50,000), and High (Above 50,000) [11].

2.2 Data Analysis

Two steps of preprocessing were applied to the collected data before the analysis: (1) *Standardization*: Standardization, transforming raw scores to z-scores, was applied to all the numerical attributes. There were 28 numerical attributes in total; (2) *One-hot encoding*: One-hot encoding techniques were applied to all the nominal attributes. There were 2 nominal attributes.

Three attribute selection methods were applied and compared, including filter methods, stepwise wrapper methods, and naturally inspired algorithms. The filter methods applied in this study included five algorithms: (1) OneR algorithm,

(2) Relief-based selection, (3) Chi-square selection, (4) Gain-ratio-based selection, and (5) Information-gain-based selection.

Both stepwise wrapper methods and naturally inspired algorithms need to have an evaluation function to work. Logistic regression was chosen as the evaluation function of both for stability and efficiency. The stepwise wrapper methods included forward and backward selection. Forward selection starts with no attributes in the model, and tests the addition of each attribute using certain comparison criteria. Backward selection starts with all candidate attributes, and tests deletion of each attribute using certain criteria. Only forward selection was used in this study.

The naturally inspired algorithm implemented was the Genetic Algorithm. Genetic Algorithm is a computational algorithm with origins in the field of biology. The tools that Genetic Algorithm uses have marks of genetic systems, including generation selection, crossover, and mutation [12]. We implemented the simple form of Genetic Algorithm described by Goldberg [13].

Weighted average F1-score was chosen as the primary evaluation criterion, because there exists an imbalance in the four income classes. A classifier that primarily guesses based on the majority class would achieve a small advantage in accuracy, but would perform worse in terms of the F1-score. Also, classification accuracy rate was used as the secondary evaluation criterion. Ten-fold cross validation was used for the estimation of both F1-score and accuracy rate.

3 Results

Five filter methods, including (1) OneR algorithm, (2) Relief-based selection, (3) Chi-square selection, (4) Gain-ratio-based selection, and (5) Information-gain-based selection, were applied to the attribute selection. The 10-fold cross validation scheme was implemented in Weka [14]. As opposed to the cross-validation in prediction or classification, no training or testing is involved in the cross-validation scheme of attribute selection. Under such a scheme, the dataset was randomly sectioned into 10 folds, and only 9 folds were used for subset attribute selection in each round. There were 10 rounds in total. The 10 selection results were summarized afterwards. The attributes selected by at least three out of the five methods (60%) were selected, yielding 14 selected attributes in total. The arithmetic mean of each attribute's ordinal ranking across all selection methods was also calculated, to enable measuring of attribute usefulness. For each single-attribute evaluator, the output of Weka showed the average merit and average rank of each attribute over the 10 folds (see Table 1).

Same as the implementation of filter methods, 10-fold cross validation scheme in Weka was used for more stable estimates. Attributes selected by at least six out of ten folds (60%) were selected, yielding 9 selected attributes in total. The selected attributes are presented in Table 2.

In alignment with the prior two attribute selection approaches, 10-fold cross validation scheme in Weka was used. Attributes selected by at least six out of ten folds (60%) were selected, yielding 22 selected attributes in total (see Table 3).

Table 1. Selected attributes subset using filter methods

Attributes	Votes*/ Average rank*	Attributes	Votes*/ Average rank*
% of Population from Students' Zip Codes over 25 with a Professional Degree	5/2.88	Admission Rate	5/12.42
Average Faculty Salary	5/3.50	Instructional Expenditure per Student	4/7.25
Average SAT Score	5/5.22	% of Students Whose Parents Have Post-High School Degree	4/9.23
Degree Completion Rate	5/6.10	Out-of-State Tuition Fee	4/10.18
% of Asian Students	5/7.22	% of Students Whose Parents were 1st Generation College Student	4/10.33
% of Students Whose Parents Have a High School Degree	5/8.58	% of 1st Gen. College Students	4/10.63
In-State Tuition Fee	5/10.88	% of Students whose Family Income classified Very High	4/11.30

**Votes Column: The number of filter methods that selected the corresponding attributes; Average Rank Column: The averaged rank values among the filter methods that selected the corresponding attributes.*

Table 2. Selected attribute subset using forward selection

Attributes	Votes*	Attributes	Votes*
Predominant Degree Type	90%	% of Students whose Parents were 1st Generation College Student	60%
Ratio between Part-time and Full-time Students	100%	% of the Population from Students' Zip Codes over 25 with a Professional Degree	100%
Degree Completion Rate	100%	% of Female Students	100%
Admission Rate	100%	Average Age of Entering College	100%
% of Asian Students	100%		

**Votes Column: The percentage of folds that selected the corresponding attributes.*

The Genetic Algorithm (GA) was the third option for attribute selection. The settings of the GA were as follows:

- Population size: 500
- Fitness function: Classification accuracy derived from Logistic Regression
- Selection Method: Tournament selection
- Crossover Type: Two-point crossover
- Crossover Rate: 0.6
- Mutation Rate: 0.03
- Stopping Criteria: 60 generations.

Table 3. Selected attributes subset using genetic algorithm

Attributes	Votes*	Attributes	Votes*
School Type	60%	% of Asian Students	100%
Predominant Degree Type	70%	% of Hispanic Students	100%
Student Size	100%	% of Students whose Family Income classified Higher Middle	80%
Instructional Expenditure per Student	90%	% of Students whose Family Income Classified Very High	100%
Ratio between Part-time and Full-time Students	100%	% of Students whose Parents have a Middle School Degree	70%
Degree Completion Rate	100%	% of Students whose Parents have a Post-High-School Degree	60%
Admission Rate	100%	% of Population from Students' Zip Codes over 25 with a Professional Degree	100%
Average SAT Score	90%	% of Female Students	100%
Out-of-State Tuition	100%	% of 1st Generation Students	60%
% of White Students	90%	Average Age of Entering College	100%
% of Black Students	60%	Average Debt	70%

**Votes Column: The percentage of folds that selected the corresponding attributes.*

Logistic Regression and Support Vector Machine with Pearson VII function kernel were used to compare the performance of the three selected attribute subsets. Ten-fold cross validation was used to estimate the classification accuracy for each classification method (see Tables 4 and 5 for individual classification results). As the most selective feature selection method (*9 attributes selected*), Forward Selection achieved acceptable F-measure. Although less selective (*22 attributes selected*), Genetic Algorithm outperformed the other two methods by both F-measure and accuracy.

Table 4. Comparisons among three selected attribute subsets using logistic regression

Attribute numbers	Accuracy	Weighted average		
		Precision	Recall	F-measure
Filter methods (N = 13)	0.691	0.688	0.691	0.686
Forward selection (N = 9)	0.736	0.733	0.736	0.731
Genetic algorithm (N = 22)	0.746	0.746	0.746	0.745

Table 5. Comparisons among three selected attribute subsets using support vector machine with Pearson VII function kernel.

Attribute numbers	Accuracy	Weighted average		
		Precision	Recall	F-measure
Filter methods (N = 13)	0.708	0.697	0.708	0.701
Forward selection (N = 9)	0.733	0.723	0.733	0.726
Genetic algorithm (N = 22)	0.755	0.745	0.755	0.747

4 Conclusion

Using College Scorecard data [10], we selected the most important factors predicting the 6-year post-graduation income of college students who used financial aid during their time at college. We compared three attribute selection methods: filter methods, forward selection, and Genetic Algorithm, in terms of classification accuracy on students' post-graduation income. We found that the attribute subset selected by the Genetic Algorithm outperformed the other two subsets when using logistic regression and support vector machine as the classification algorithm.

We wish to draw attention to how higher numbers of students in a cohort who grew up in Zip code areas where over 25% of the population hold a Professional Degree was predictive of more college graduates likely being classified as High income. This finding is aligned with evidence about how geography or "where you grow up" impacts life outcomes. Chetty et al. [15] identified that areas with lower racial segregation and income inequality, but higher social capital[1] and family stability are associated with greater opportunities for intergenerational social mobility. In the current research, the role of geography for post-graduation incomes in the case of neighborhood Professional Degree attainment signifies social stratification in graduate labor markets. The finding may stem from unequal access to support for education and careers. This would reinforce the Effectively Maintained Inequality model that predicts that as access to education widens, higher socio-economic status students will seek "horizontal differentiation" by accessing *qualitatively* distinctive or superior types of education that maintain their advantage in society [17, 18].

We are *not* arguing that young people from disadvantaged neighborhoods should not attend higher education. Attaining a Bachelor's degree remains an excellent "investment" to enhance career prospects. Yet our findings showing a disparity of post-graduation income according to "where you grow up" suggests a need for greater support for students *both* in college access and in transitions to the labor market, especially given rising tuition fees and associated concerns about student debt [19].

Appendix A

The dataset analyzed in this study can be accessed at https://collegescorecard.ed.gov/data/.

30 potential attributes include:

Group One: School information

1. School Type (e.g. private school)
2. Predominant Awarded Degrees (e.g., bachelor degree)
3. Student Size
4. Instructional Expenditure per Student

[1] Social capital represents trust, solidarity, and reciprocity in collective social interactions and engagement in community-based activities [16].

5. Ratio between Part-time and Full-time Students
6. Degree Completion Rate
7. Average Faculty Salary

Group Two: Admission information

8. Admission Rate
9. Average SAT Score

Group Three: Cost information

10. In-State Tuition
11. Out-of-State Tuition

Group Four: Student information

12. Percentage of White Students
13. Percentage of Black Students
14. Percentage of Asian Students
15. Percentage of American Indian Students
16. Percentage of Hispanic Students
17. Percentage of Female Students
18. Percentage of First-Generation Students
19. Average Age of Entering College
20. Average Debt

Group Five: Family and community information

21. Percentage of Students whose Family Income was classified as Low
22. Percentage of Students whose Family Income was classified as Lower Middle
23. Percentage of Students whose Family Income was classified as Higher Middle
24. Percentage of Students whose Family Income was classified as High
25. Percentage of Students whose Family Income was classified as Very High
26. Percentage of Students whose Parents were 1st Generation College Student
27. Percentage of Students whose Parents Have a Middle School Degree
28. Percentage of Students whose Parents Have a High School Degree
29. Percentage of Students whose Parents Have a Post-High-School Degree
30. Population from Students' Zip Codes over 25% with a Professional Degree.

References

1. Autor, D.H.: Skills, education, and the rise of earnings inequality among the 'other 99 percent'. Science **344**(6186), 843–851 (2014)
2. Chetty, R., Friedman, J., Saez, E., Turner, N., Yagan, D.: Mobility report cards: the role of colleges in intergenerational mobility. Technical report, Stanford University (2017)
3. Hout, M.: Social and economic returns to college education in the United States. Ann. Rev. Sociol. **38**, 379–400 (2012)

4. National Center for Educational Statistics [NCES]. Percentage of 18- to 24-year-olds enrolled in degree-granting postsecondary institutions, by level of institution and sex and race/ethnicity of student: 1970 through 2015. http://nces.ed.gov/programs/digest/d15/tables/dt15_302.60.asp?current=yes. Accessed 1 Mar 2018
5. Beaudry, P., Green, D.A., Sand, B.M.: The declining fortunes of the young since 2000. Am. Econ. Rev. **104**(5), 381–386 (2014)
6. Valletta, R.G.: Recent flattening in the higher education wage premium: polarization, skill downgrading, or both? In: Education, Skills, and Technical Change: Implications for Future US GDP Growth. University of Chicago Press (2017)
7. Federal Reserve Bank of New York. The Labor Market for Recent College Graduates. https://www.newyorkfed.org/research/college-labor-market/index.html. Accessed 1 Mar 2018
8. Altonji, J.G., Arcidiacono, P., Maurel, A.: The analysis of field choice in college and graduate school: determinants and wage effects (no. w21655). National Bureau of Economic Research (2015)
9. Witteveen, D., Attewell, P.: The earnings payoff from attending a selective college. Soc. Sci. Res. **66**, 154–169 (2017)
10. U.S. Department of Education. https://www.newyorkfed.org/research/college-labor-market/index.html. Accessed 1 Mar 2018
11. U.S. Census Bureau: Distribution of Personal Income 2010 (2010). https://www.census.gov/2010census/data/. Accessed 1 Mar 2018
12. Beasley, J.E., Chu, P.C.: A genetic algorithm for the set covering problem. Eur. J. Oper. Res. **94**(2), 392–404 (1996)
13. Goldberg, D.: Genetic Algorithms in Optimization, Search and Machine Learning. Addison-Wesley, Reading (1989)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
15. Chetty, R., Hendren, N., Kline, P., Saez, E.: Where is the land of opportunity? The geography of intergenerational mobility in the United States. Q. J. Econ. **129**(4), 1553–1623 (2014)
16. Putnam, R.D.: Our Kids: The American Dream in Crisis. Simon and Schuster, New York (2016)
17. Lucas, S.R.: Effectively maintained inequality: education transitions, track mobility, and social background effects. Am. J. Sociol. **106**, 1642–1690 (2001)
18. Lucas, S.R., Byrne, D.: Effectively maintained inequality in education: an introduction. Am. Behav. Sci. **61**(1), 3–7 (2017)
19. Avery, C., Turner, S.: Student loans: do college students borrow too much—or not enough? J. Econ. Perspect. **26**(1), 165–192 (2012)