FEATURE SELECTION AND CLASSIFICATION OF POST-GRADUATION INCOME OF COLLEGE STUDENTS IN THE UNITED STATES

by

QIANG HAO

(Under the Direction of Khaled Rasheed)

ABSTRACT

This study investigated the most important attributes of the 6-year post-graduation income of college graduates who used financial aid during college time 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. Five groups of machine learning algorithms and three ensemble learning methods were applied to the dataset for classification based on the best selected attribute subsets. The common attributes selected by at least two selection methods were further discussed.

INDEX WORDS: machine learning; post-graduation income classification; ensemble learning; College Scorecard

FEATURE SELECTION AND CLASSIFICATIONOF POST-GRADUATION INCOME OF COLLEGE STUDENTS IN UNITED STATES

by

QIANG HAO

M.S., The University of Hong Kong, Hong Kong, 2012

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2017

FEATURE SELECTION AND CLASSIFICATIONOF POST-GRADUATION INCOME OF COLLEGE STUDENTS IN UNITED STATES

by

QIANG HAO

Major Professor:	Khaled Rasheed
Committee:	Hamid Reza Arabnia
	Yi Hong

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. INTRODUCTION

Higher education has expanded on a mass and global scale since the late 20th century. The great majority of countries around the world have experienced both absolute growth in the total number and relative growth in the proportion of the age cohort enrolled in higher education institutions (Marginson 2016; Schofer & Meyer 2005). Being the forerunner in the massification of higher education from the 1960s, the United States has been no exception to this trend. Nonetheless, recent decades have witnessed stagnating enrolment growth at four-year institutions and persistent under-representation among particular societal groups. This has been raised as a major concern in the context of a "rising skill premium" (Autor 2014 p. 2) owing to skill biased technological change and the globalisation of production. Studies have consistently identified that on average graduates significantly outperform those with only a high-school diploma in the US labour market. Following this logic, higher education is viewed as crucial to equipping individuals from diverse backgrounds to pursue a career in modern labour markets and to drive national economic growth, particularly in the fields of labour economics (Becker 1964; Goldin & Katz 2009) and sociology (Blau & Duncan 1967; Hout 2012)

However, higher education has become an expensive endeavour and young people are increasingly expected to self-finance their studies, especially through student loans. This is important given growing indications that the returns to higher education are not spread evenly across graduate populations. Put another way, as the individual costs of higher education have risen, graduates have become a progressively more heterogeneous

group with diverse outcomes in the labour market. While some graduates receive a substantial premium in salaries, others encounter unemployment, under-employment, and low-salaries (Beaudry et al. 2014; Federal Reserve Bank of New York 2016; Vedder 2013). As an explanation, research has shown that the choice of major and institution attended are important factors in determining the salaries received by graduates. Yet, there is a pressing need to better understand how other factors contribute to the increasingly varied outcomes of graduates in the US labour market. In response, this study explored the most important factors of 6-year post-graduation income of college graduates who used student aid from the U.S. Department of Education during their time in college, and to what extent of accuracy the select factors can be used to classify post-graduation income of college graduates.

2. LITERATURE REVIEW

2.1 The Graduate Premium and Student Debt

The proportion of 18 to 24 year olds enrolled at degree granting higher education institutions in the United State increased steadily from 25.7 percent in 1970 to 32.0 percent by 1990 and then to 40.0 percent in 2014 (NCES, 2016a). In recent years, however, growth at four-year higher education has stagnated. In the decade between 2004 and 2014, enrollments at 4-year institutions increased by only 0.8 percent points to reach 29.4 percent of young people (NCES, 2016b). Moreover, participation in higher education in the US remains unequally distributed across society with regards to socio-economic status, ethnicity, and gender. First, 80.7 of high school graduates from the top 20 percent of family income levels were enrolled in college in 2012, compared to only 50.9 percent of those from the bottom 20 of family income levels (NCES 2016b). Second, according to ethnicity, college enrolments among high school graduates stood at 81.5 percent (Asian), 70.3 percent (Hispanic) 65.7 percent (White), and 56.4 percent (Black) (NCES 2016c). Third, with regard to gender, 71.3 percent of female high school graduates enrolled at college, compared to only 61.3 percent of males (NCES 2016d).

Furthering enrolment growth and tackling under-representation among certain groups in higher education is viewed as a core mechanism to promoting both national and individual level prosperity, in equal respects. This reflects a “rising skill premium” (Autor 2014 p. 2) in the US labour market. On the one hand, recent technological advancements are viewed as complementing the productivity of highly-educated workers

while simultaneously automating many routine jobs of lowly-educated workers (Brynjolfsson & McAfee 2014; Goldin & Katz 2009). On the other hand, the globalisation of production has resulted in the relocation of jobs in many industries exposed to import competition. An implication is that a significant number of jobs have been offshored since the late 20th century, especially in terms of manufacturing jobs being relocated to China (Autor et al. 2016; Blinder & Krueger 2009).

Higher education is, thus, generally viewed as an excellent "investment" that should be encouraged by families, schools, communities, and policy makers (Autor 2014; Goldin & Katz 2009; Hout 2012). Indeed, the returns of a college degree vis-à-vis a high school degree or the "graduate premium" has expanded considerably in recent decades. For example, Autor (2014 p. 847) demonstrates that the "graduate premium" doubled in real terms between 1979 and 2012. That is, the gap of 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. In addition to economic returns, Hout (2012) contends that higher education "makes life better" (p. 394) through a host of social benefits in the realms of community relations, health, family stability, and social connections.

Nevertheless, a corresponding trend has been a rapid increase in the costs of higher education (Avery & Turner, 2012; Brown et al. 2014; Houle, 2013; Wolff et al. 2014). Wolff et al. (2014 p. 2) estimate that higher education tuition fees have risen by 250 percent since the early 1980s when measured in dollars of constant purchasing power. The increases in cost have not been met by a corresponding increase in grants or other forms of aid, especially as state funding for higher education has come under pressure

(Houle, 2012). To fill this funding gap, students have become more dependent on student loans to finance higher education studies. Brown et al. (2014) demonstrate that the proportion of 25 year olds with student debt grew from 27 percent in 2004 to 43 percent in 2012 (p. 7), while during the same period the average debt being held grew by 70 percent to $25,000 (p. 5). Across the US, the total amount of student debt reached $966 billion in 2012, representing a three-fold increase since 2004 (p. 3).

2.2 Graduates as a Heterogeneous Population

Student loans can be viewed as a "bridge" to enable all students, irrespective of their financial status, to access the rewards of a higher education (Becker & Tomes, 1994). Yet, taking on debt to finance higher education inevitably carries a degree of risk. This point is especially pertinent given growing evidence of divergence in the labour market returns of higher education. In other words, as the number of college graduates has expanded, college graduates have become an increasing heterogeneous population with increasingly disparate labour market outcomes (Beaudry 2014; Federal Reserve Bank of New York 2016; Vedder 2013). Data from the Federal Reserve Bank of New York (2016) puts unemployment among college graduates aged between 22 and 27 at 4.6 percent. More significantly, the same dataset shows that 44.9 percent of these recent graduates were under-employed or employed in a job that "typically does not require a college degree". This latter point is reflected in graduate earnings with the dataset also illustrating that 13.8 of recent graduates are employed in "low-wage jobs" that tend to pay below $25,000 per annum. Reflecting this trend, Haughwout et al. (2015) identified that half of college graduates from the 2009 cohort who took student loans have defaulted, gone delinquent, or made no progress in paying back their debt.

To better understand who benefits the most and least from higher education, there is a pressing need to identify the main factors that contribute to the relative earnings of different types of college graduates. The literature to date has focused on institutional factors and course related factors, while there is a limited but emerging body of research looking at the role of student characteristics (i.e., socio-economic background, ethnicity, and gender). A first factor is that the type of higher education institution attended has a sizable bearing on the economic returns of higher education. There is strong evidence that students enrolling at community colleges, for-profit institutions, and online institutions have lower completion rates and are less likely to receive a salary premium upon graduation (Avery & Turner 2012; Cellini & Chaudhary 2014; Deming et al. 2016). For example, Deming et al. (2016) found that college graduates of online for-profit institutions were 22 percent less likely to receive a call-back from employers. Conversely, other studies have identified highest returns among college graduates of highly selective or flagship institutions (Dale & Krueger 2011; Eide et al. 2016; Hoekstra 2009). Hoekstra (2009) identified a 24 percent earnings premium for graduating from a flagship state university, while Dale and Krueger (2011) note a large earning premium for selective colleges for graduates from less-educated families and among black and Hispanic graduates.

A second set of factors are related to course of study including choice of major and academic performance. There is growing evidence that graduates of majors associated with science, technology, engineering, and maths (STEM) enjoy higher salaries and lower incidents of unemployment, compared to graduates from non-technical majors (Altonji et al. 2015; Federal Reserve Bank of New York 2016; Kim et al. 2015). Indeed,

Kim et al. (2015) estimate that lifetime median earnings gains for STEM degree holders (US$800,000) are more than double the relative gains for a social science degree (US$374,000). In addition, data from the Federal Reserve Bank of New York (2016) demonstrates that the highest six earning majors for recent graduates were all variants of engineering (i.e., chemical, electrical, mechanical, industrial, computer, and aerospace). Further studies have identified that academic performance can also be an important factor for determining success in the labour market. More specifically, high school GPA (French et al. 2015) was shown to be a strong predictor of future earnings, while university GPA was significant for students graduating from less selective institutions.

While less researched, a third set of factors are associated with the characteristics of individual college graduates. A limited body of research has identified varying labour market returns of higher education according to socio-economic background (Brown et al. 2014; Herbstein 2016; Mettler 2014). Bartik and Hershbein (2016) identified that college graduates from families with incomes above 185 percent the poverty line had career earnings 135 percent greater than high school diploma holders, compared to only being 69 percent greater for college graduates from families with incomes below 185 percent the poverty line. Further disparities in labour market outcomes can be observed in terms of ethnicity (Andrews et al. 2016; Jones & Schmitt, 2014; Kroeger et al. 2016). Jones and Schmitt (2014) note that young black college graduates are disproportionately prone to unemployment and under-employment, which stood at 12.4 percent and 55.9 percent in 2013, respectively. Also, while women have surpassed men in higher education enrolment rates a gender wage gap remains persistent (DiPrete & Buchmann, 2013; Kroeger et al., 2016; Thornton & McDonald, 2015). Between 2000 and 2015

incomes for young male college graduates increased by 8.1 percent, while declining by 6.8 percent for young female college graduates over the same period (Kroeger et al. 2016).

3. RESEARCH QUESTIONS

The research questions that guided this study include:

1. What are the most important attributes of post-graduation income of college students who graduate with debt repayment obligations?
2. To what extent can the selected attributes classify post-graduation income of college students who graduate with debt repayment obligations?

4. RESEARCH DESIGN

4.1 Data Collection

The data for this study was the latest release in October, 2015 by College ScoreCard under the United States Department of Education (https://collegescorecard.ed.gov/data/). 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 range from 1996 to 2013, but the 6-year post-graduation income data are only available for the year 1997, 1999, 2001, 2003 and 2005. The selected target of this project is the mean value of the 6-year post-graduation income of a student cohort. Attributes were filtered based on domain knowledge firstly. The factors deemed as less relevant were excluded, such as latitude of the institution, accreditor of the institution, or percent of students who passed away within 6 years after graduation. Thirty potential attributes under five groups are included in this project. The groups are the following:

1. School
2. Admission
3. Cost
4. Student Cohort
5. Socioeconomic Status of Students' Family

Some attributes in certain groups are not available before 2000, such as admission rate in the Admission Group. Therefore, only three years' data, including 2001, 2003, and 2005 were used in this study. 1429 cohorts were included for the data analysis.

The target, mean income value of each cohort, was discretized into four groups based on the information from the American Individual Income Distribution (U.S. Census Bureau, 2010).

- Very low: From 0 to 25000
- Low: From 25000 to 37500
- Middle: From 37500 to 50000
- High: Above 50000

Two steps of preprocessing were applied to the collected data before the analysis:

1. Standardization: Z-score standardization was applied to all the numerical attributes. There were 28 numerical attributes in total. Given a feature vector x_j, the z-score standardization formula is defined as:

$$z_j = (x_j - \textit{Mean of } x_j) / \textit{Standard deviation of } x_j$$

2. One-hot encoding: One-hot encoding techniques were applied to all the nominal attributes. There were 2 nominal attributes in total. By one-hot encoding, we mean the technique that transforms a multi-categorical feature into a set of binary features that correspond to each of the original categories.

4.2 Data Analysis

The data analysis of this project had two phases. The first phase was attribute selection. 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 include the following five algorithms:

- OneR algorithm: The OneR algorithm generates one rule for each feature in the data, then selects the rule with the smallest total error as its one rule. (Witten, et al., 2016).
- Relief-based selection: Relief-based selection weighs each feature according to its relevance to a class. All weights are set to zero initially, and updated iteratively later. In each iteration, a random instance i in the dataset would be chosen and estimates would be given on how well each feature value of this instance distinguishes among instances close to i (Kira & Rendell, 1992).
- Chi-square selection: Chi-square selection calculates Chi-square statistics between each feature and the target class, and keeps the features that have higher Chi-square statistics (Witten, et al., 2016). Chi-square statistics measure the dependency between two given variables.
- Information-gain-based selection: Information-gain-based selection performs selection based on the calculation of information gain. Information Gain is the expected reduction in entropy caused by partitioning the data according to a given attribute (Azhagusundari & Thanamani, 2013).
- Gain-ratio-based selection: Gain-ratio-based selection performs selection based on information gain-ratio of each feature. In comparison with information-gain-based selection, information gain ratio reduces the bias towards multi-valued attributes by considering both the number and size of branches (Witten, et al., 2016).

Both stepwise wrapper methods and naturally inspired algorithms need to have an evaluation function to work. Logistic regression was chosen as the evaluation function for both of them for the consideration of stability and efficiency. The stepwise wrapper methods include forward and backward selection. Forward selection starts with no attributes in the model, and tests the addition of each attribute using certain comparison criteria. In contrast, 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 in this study is the Genetic Algorithm (GA). GA is a computational algorithm that has origins in biology. The major tools that GA uses have marks of genetic systems, including generation selection, crossover, and mutation (Beasley & Chu, 1996). The simple GA described by Golberg (1989) was implemented in this study.

The second phase of this project was the exploration of the extent to which the selected attributes can classify post-graduation income of college students, and how they perform compared to the whole attribute set. Ten single machine learning algorithms in five groups were applied to the dataset for classification purposes. The five groups of algorithms include:

1. *Bayes-based algorithms*:
    - Naive Bayes Update: Naive Bayes Update operates in the same way as Naïve Bayes but uses Kernel density estimation instead of normal density measures for numeric attributes (Gu, et al., 2009).

- Bayes Net: Bayes Net represents a set of random variables and their conditional dependencies via a directed acyclic graph, and can use such conditional relationships for classification purposes.

2. *Function-based algorithms*:
    - Logistic Regression: Logistic regression is a regression model where the target is categorical. Specifically, multinomial logistic regression was used in this study. Multinomial logistic regression extends logistic regression to multiclass problems.
    - Support Vector Machine (SVM): SVM is a supervised learning algorithm that analyzes data for classification by learning a hyperplane that achieves the largest separation among different classes (Hsu & Lin, 2002).
    - Multilayer Perceptron: Multilayer perceptron is an artificial neural network that utilizes a technique called backpropagation for classification purposes. Multilayer perceptron extends the capability of standard linear perceptron so that it can distinguish data that are not linearly separable (Cybenko, 1992).
3. *Instance-based algorithms*:
    - Distance-weighted K-Nearest Neighbor (KNN): KNN is a non-parametric method that can be used for classification purposes, by taking a majority vote of the K nearest neighbors, and assigning their most common class to the object (Altman, 1992).
4. *Tree-based algorithms*:

- J48: J48, as a simple decision tree classifier, creates a decision tree by always using the attributes that discriminate the data most clearly first.
- Multiclass Alternating Decision Tree (ADTree): ADTree is a classification technique that combines decision trees with the predictive accuracy of boosting into a set of interpretable classification rules (Holmes, et. al., 2002).

5. *Rule-based algorithms*:
    - OneR: In addition to feature selection, OneR algorithm can also be used for classification purposes. When OneR is used for classification purposes, it simply uses the rule with the smallest total errors for classification. (Witten, et al., 2016).
    - JRIP: JRip examines data in different classes in increasing size and generates an initial set of rules using incremental reduced error. After that, JRIP proceeds by treating all the examples of a particular judgment in them as a class, and generates a set of rules that cover all the examples of that class (Rajput, et. al., 2011).

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. In addition, 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.

5. RESULTS

5.1 Attribute Selection

5.1.1 Attribute Selection using Filter Methods

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 (Hall, et al., 2009). 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.

Table 1

Selected Attributes Subset using Filter Methods.

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

5.1.2 Attribute Selection using Forward Selection

Same as the implementation of filter methods, the 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.

Table 2

Selected Attribute Subset using Forward Selection.

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

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

5.1.3 Attribute Selection using Genetic Algorithm

The Genetic Algorithm (GA) was as the third option for attribute selection. The settings of the GA in this study 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

Same as the other 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. The selected attributes are summarized in Table 3.

Table 3

Selected Attributes Subset using Genetic Algorithm.

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

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

5.1.4 Comparisons among the three selected attribute subsets

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. The individual classification results are presented in Table 4 and Table 5. The attribute subset selected by the Genetic Algorithm outperformed the other two using both classification algorithms.

Table 4

Comparisons among Three Selected Attribute Subsets Using Logistic Regression.

Logistic Regression	Accuracy	Weighted Average		
		Precision	Recall	F-measure
Attribute Subset Selected by Filter Methods (N = 13)	0.691	0.688	0.691	0.686
Attribute Subset Selected by Forward Selection (N = 9)	0.736	0.733	0.736	0.731
Attribute Subset Selected by 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.

Support Vector Machine with Pearson VII function kernel	Accuracy	Weighted Average		
		Precision	Recall	F-measure
Attribute Subset Selected by Filter Methods (N = 13)	0.708	0.697	0.708	0.701
Attribute Subset Selected by Forward Selection (N = 9)	0.733	0.723	0.733	0.726
Attribute Subset Selected by Genetic Algorithm (N = 22)	0.755	0.745	0.755	0.747

5.2 Classification Using the Best Attribute Subset

5.2.1 Classification using Single Learners

Ten single machine learning algorithms from five groups were applied to the best attribute subset, which is selected by the Genetic Algorithm (see Table 3). The five groups of algorithms include:

1. *Bayes-based algorithms*: Naive Bayes Update, Bayes Net
2. *Function-based algorithms*: Logistic Regression, Support Vector Machine (SVM) with kernel as Pearson VII function, Multilayer Perceptron with one hidden layer and 13 neurons
3. *Instance-based algorithms*: Distance-weighted K-Nearest Neighbor (KNN, weight = 1/distance)
4. *Tree-based algorithms*: J48, Multiclass Alternating Decision Tree (ADTree)
5. *Rule-based algorithms*: OneR, JRIP

The 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 F1-score. In addition, classification accuracy rate was used as the secondary evaluation criterion. Ten-fold cross validation was used for the estimate of both F1-score and accuracy rate. The results from each algorithm group are presented in Figure 1.

The top three performers were identified as Support Vector Machine, and K-Nearest Neighbor with K equal to 1 and 10 respectively, and their detailed classification results are presented in Table 6.

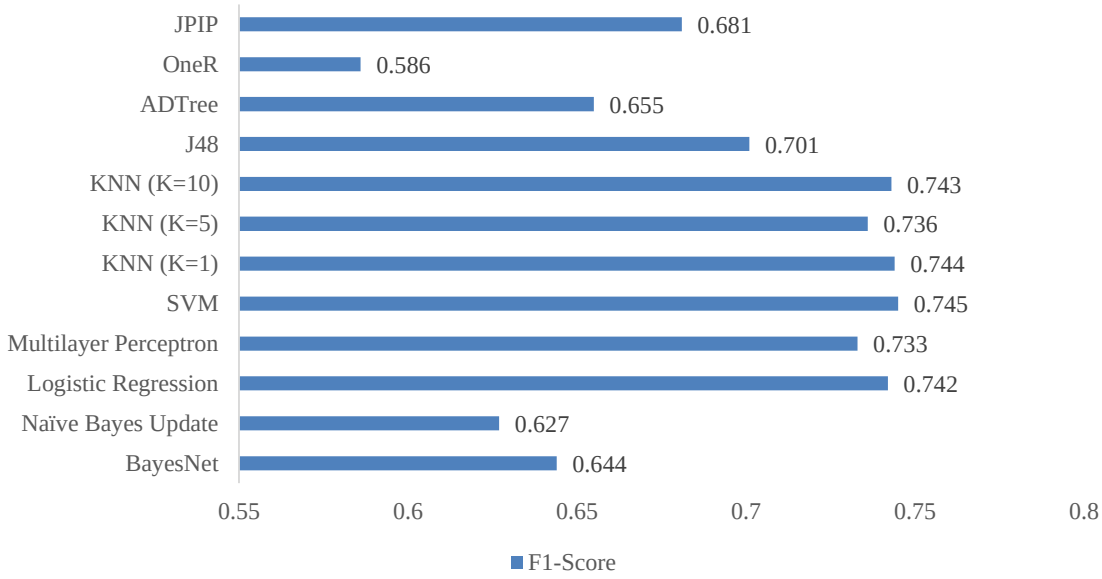

Figure 1. Performance Comparison among Single Learners Using the Best Attribute Subset.

Table 6.

Top Three Performers of Single Learners Using the Best Attribute Subset.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Support Vector Machine (*kernel = Pearson VII function*)	0.753	0.743	0.753	0.745
K-Nearest Neighbor (*distance weight = 1/distance; K = 1*)	0.745	0.744	0.745	0.744
K-Nearest Neighbor (*distance weight = 1/distance; K = 10*)	0.747	0.748	0.747	0.743

5.2.2 Classification using Ensemble Learning

Three methods of ensemble learning were explored in this study, including randomization, bagging and boosting. Bagging refers to the technique that resamples the

data with replacement randomly, so multiple models can be trained on resampled data and votes taken (Breiman, 1996). In contrast, boosting refers to the technique that allows iterative learning and lets future learners focus more on previously misclassified examples through reweighting (Breiman, 1998). Randomization refers to the technique of randomizing the algorithm instead of the training dataset (Dietterich, 2000).

Bagging and boosting were used as wrapper methods with each of the single learners from the previous section as the core function using the same configurations. Random Tree and Random Forest were applied as randomization methods. Also, Random Forest with bagging and boosting separately were explored as combination strategies.

The comparisons between randomization methods and the single learners used in previous sections are presented in Figure 2. Although the performance of Random Tree is disappointing, Random Forest outperformed all other learning algorithms.

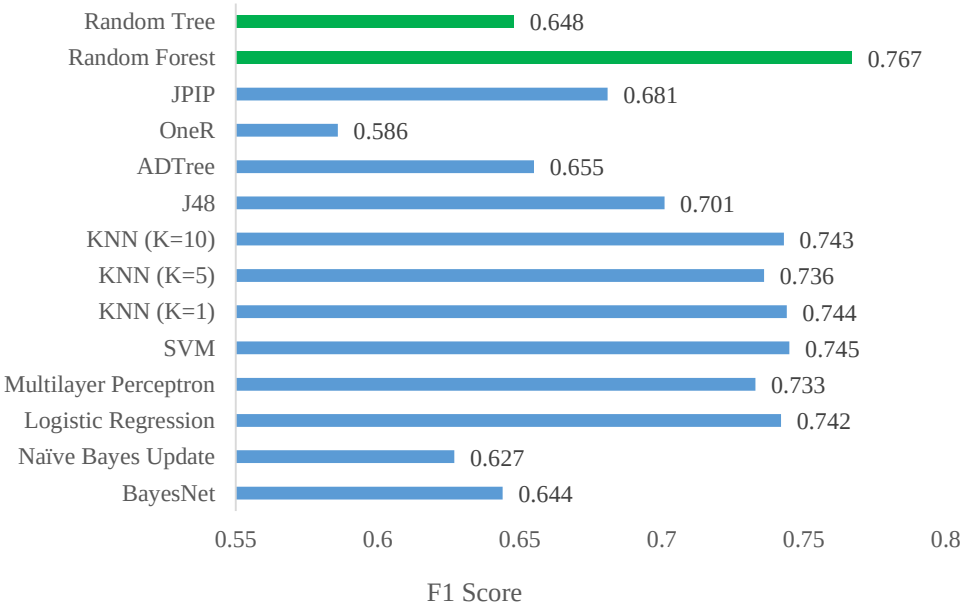

Figure 2. Performance Comparison between Randomization Methods and Single Learners Using the Best Attribute Subset.

The results of ensemble learning grouped by bagging and boosting are presented in Figure 3 and Figure 4. Both bagging and boosting techniques improved the performance of more than 60% of all algorithms to varied extent.

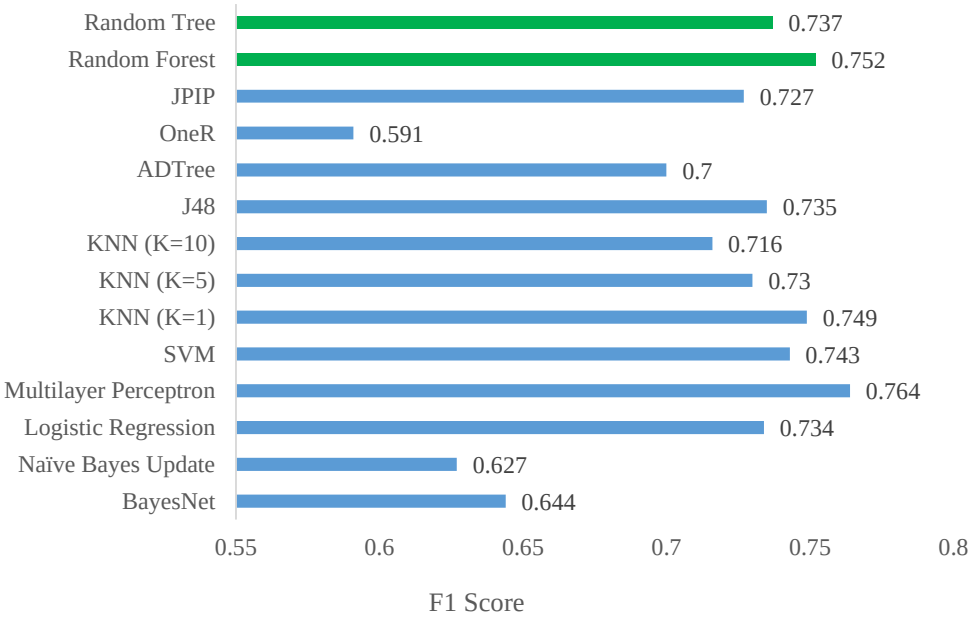

Figure 3. Using Bagging as the Wrapper Method for the Classification of Post-Graduation Income with the Best Attribute Subset.

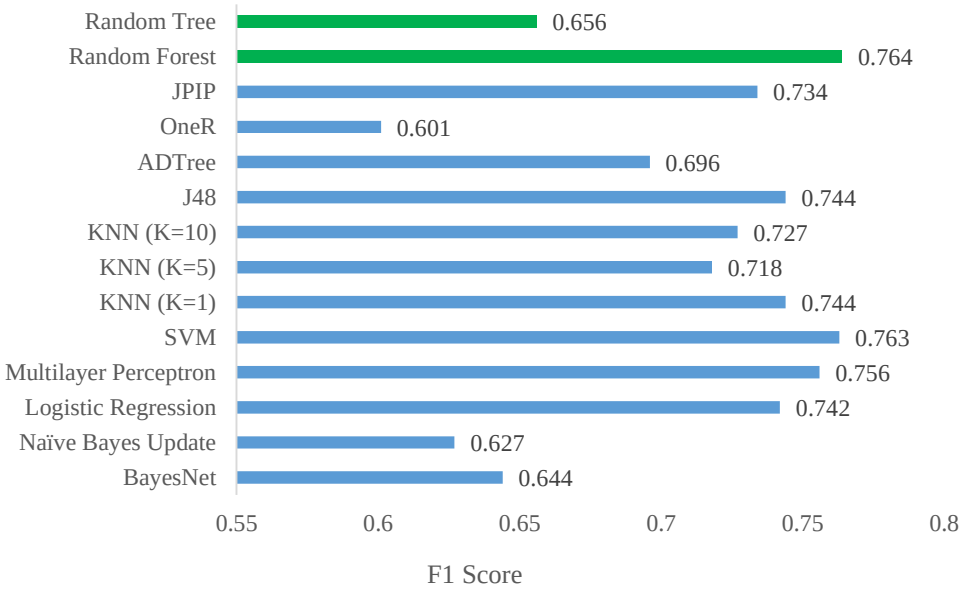

Figure 4. Using Boosting as the Wrapper Method for the Classification of Post-Graduation Income with the Best Attribute Subset.

The top three performers with ensemble learning were identified as Random Forest, Multilayer Perceptron with Bagging, and Support Vector Machine with Boosting. The details of the top three ensemble learning classifications are presented in Table 7. Random Forest performed the best on both F1-Score (0.767) and accuracy rate (0.770).

Table 7.

Top Three Performers with Ensemble Learning Using the Best Attribute Subset.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Random Forest	0.770	0.769	0.77	0.767
Multilayer Perceptron (*one hidden layer and 13 neurons*) with Bagging	0.768	0.763	0.768	0.764
Support Vector Machine (*kernel = Pearson VII function*) with Boosting	0.767	0.763	0.767	0.763

5.3 Classification Using All Attributes

5.3.1 Classification using Single Learners

For the purpose of comparison, the same classification procedure as described in the section 5.2 was applied to all 30 attributes. The performance of all single learners on the whole attribute set is presented in Figure 5.

Very similar to the results of using the best attribute subset, Multilayer Perceptron, and K-Nearest Neighbor, and Support Vector Machine were identified as the top three single learners. The details of the top three performers are presented in Table 8.

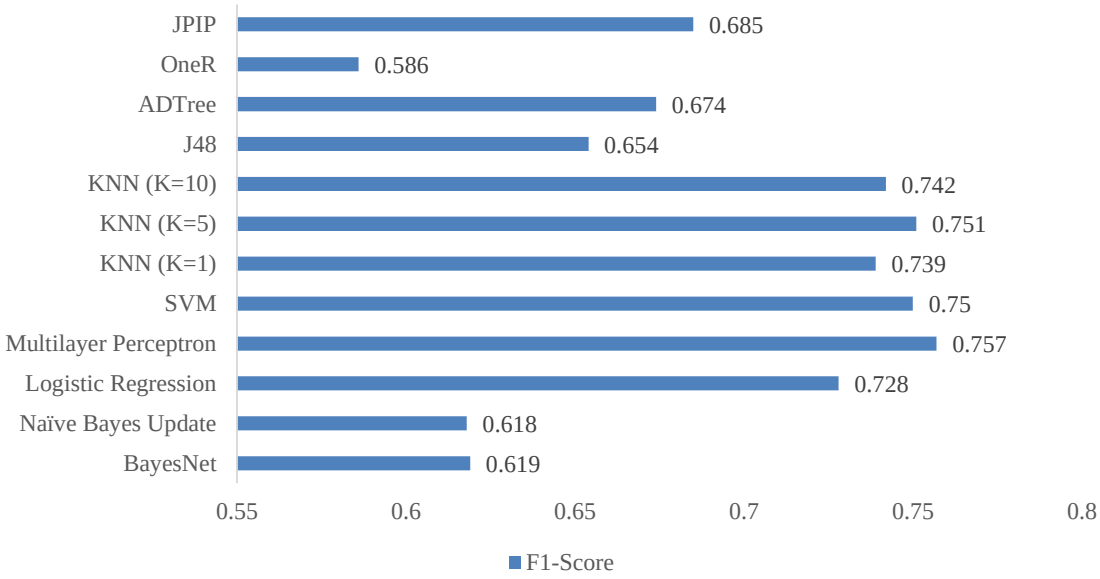

Figure 5. Performance Comparison among Single Learners Using All Attributes.

Table 8.

Top Three Performers of Single Learners Using All Attributes.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Multilayer Perceptron (*one hidden layer and 18 neurons*)	0.761	0.757	0.761	0.757
K-Nearest Neighbor (*distance weight = 1/distance; K = 5*)	0.755	0.753	0.755	0.751
Support Vector Machine (*kernel = Pearson VII function*)	0.758	0.746	0.758	0.75

5.3.2 Classification using Ensemble Learning

The comparisons between randomization methods and the single learners in previous sections are presented in Figure 6. Same as the results of using the best attribute subset, Random Forest outperformed all other learning algorithms.

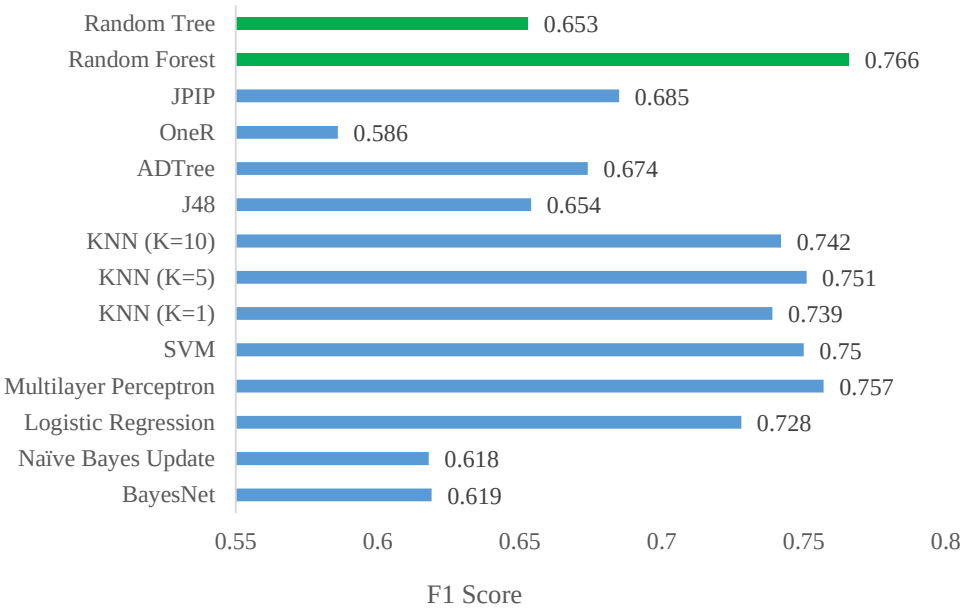

Figure 6. Performance Comparison between Randomization Methods and Single Learners Using All Attributes.

The results of ensemble learning grouped by bagging and boosting were presented in Figure 7 and Figure 8. Both bagging and boosting techniques improved the performance of more than 60% of all algorithms to varied extent.

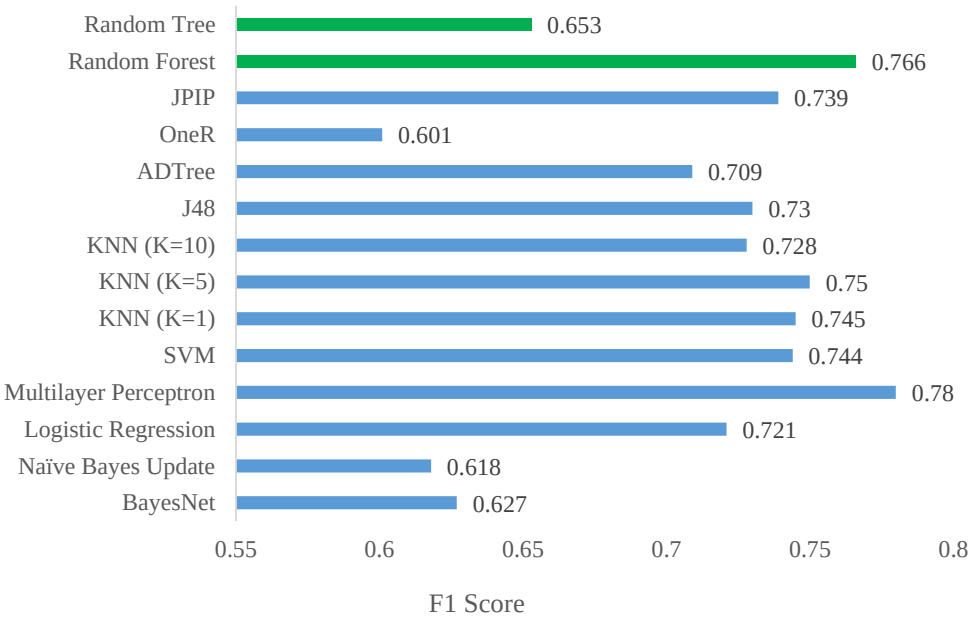

Figure 7. Using Bagging as the Wrapper Method for the Classification of Post-Graduation Income with All Attributes.

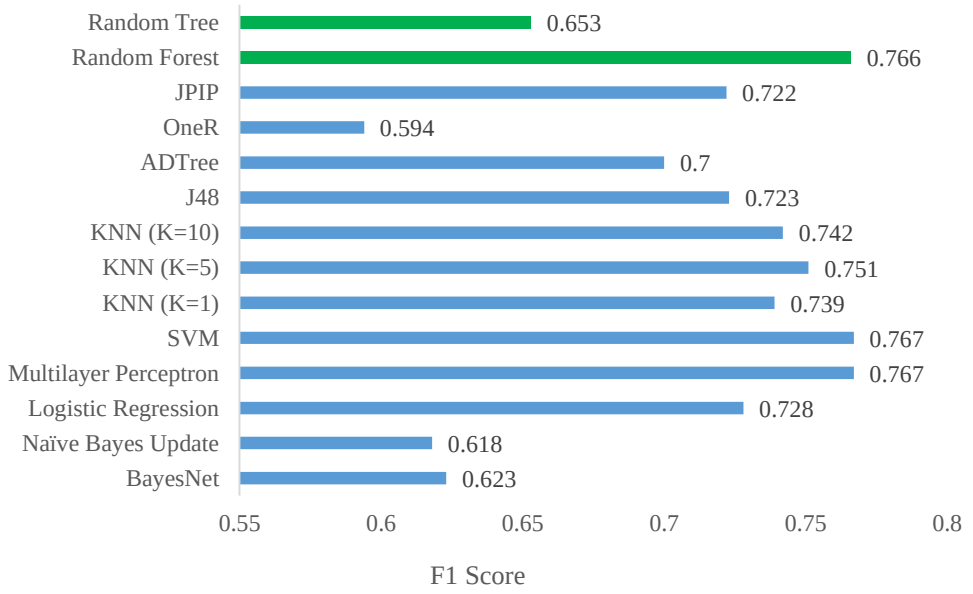

Figure 8. Using Boosting as the Wrapper Method for the Classification of Post-Graduation Income with All Attributes.

The top three performers with ensemble learning were identified as Multilayer Perceptron with Bagging, Multilayer Perceptron with Boosting, and Support Vector Machine with Boosting. The details of the top three ensemble learning classifications are presented in Table 9. Multilayer Perceptron with Bagging performed the best regarding of both F1-Score (0.78) and accuracy rate (0.784).

Table 9.

Top Three Performers with Ensemble Learning Using All Attributes.

Algorithm	Accuracy	Weighted Average		
		Precision	Recall	F1-Score
Multilayer Perceptron (*one hidden layer and 18 neurons*) with Bagging	0.784	0.782	0.784	0.78
Multilayer Perceptron (*one hidden layer and 18 neurons*) with Boosting	0.768	0.767	0.768	0.768
Support Vector Machine (*kernel = Pearson VII function*) with Boosting	0.770	0.767	0.770	0.767

6. DISCUSSIONS

Using the data from the College Scorecard (U.S. Department of Education 2016), we selected the most important factors predicting the 6-year post-graduation income of college students who used financial aid during their college time. Specifically, we compared three attribute selection methods, including filter methods, forward selection, and Genetic Algorithm, in terms of the classification accuracy on students' post-graduation income. In this process, 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.

Based on our findings, we wish to draw attention to some attributes that were selected by at least two selection methods, including socio-economic status related factors, admission rate, offered degree and SAT score.

First, regarding socio-economic status, higher numbers of students in a cohort who grew up in Zip code areas where over 25 percent of the population had a Professional Degree was predictive of more college graduates likely being classified as High income. This finding is in line with social capital theories that stress the role of local community factors such as positive role models, mutual trust, and cooperation for the socialisation and outcomes of young people (Coleman, 1988; Putnam, 1995; Putnam, 2016). In addition, this trend is highly correlated to the relationship between graduates' income and the ratio of high-income parents in the same cohort. The finding in terms of the relationship between students' SAT score and post-graduation income reinforces the

findings of prior research on post-graduation incomes (e.g. Hoekstra, 2009). It is worth noting that all the three relationships are highly correlated with each other. The parents who have higher income tend to be more capable of giving the children a better education, which leads to more competiveness in SAT, and students with higher SAT scores tend to attend more selective schools and thus are more likely to receive higher income after graduation (see Figure 9). This correlation is once again confirmed at a much larger scale using College Scoreboard data.

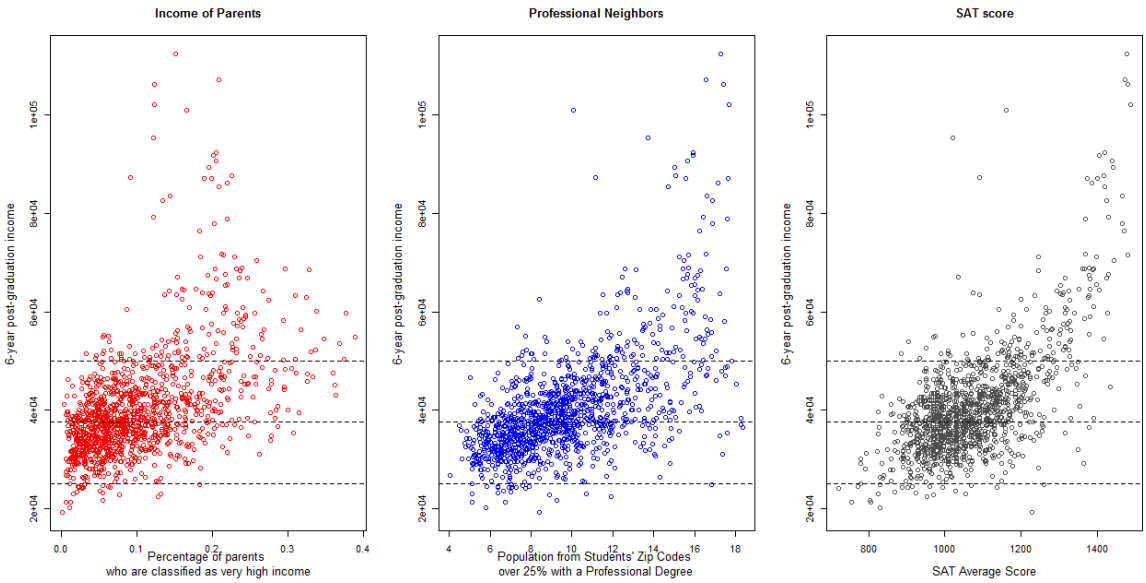

Figure 9. Relationships among six-year post-graduation income, percentage of parents who are classified as very high income, population from student's zip codes over 25% with a professional degree, and SAT score.

Secondly, as the percentage of students whose parents were a 1st generation college student increases, the post-graduation income of students is more likely to be classified as Low (see Figure 10). This finding may stem from parents with a college education being able to provide more informed support and guidance to their children than their

counterparts. Similarly, research has identified that first generation students are often handicapped by attending less selective institutions and being more likely to face difficulties with their academic studies at college (Pascarella, et al. 2004).

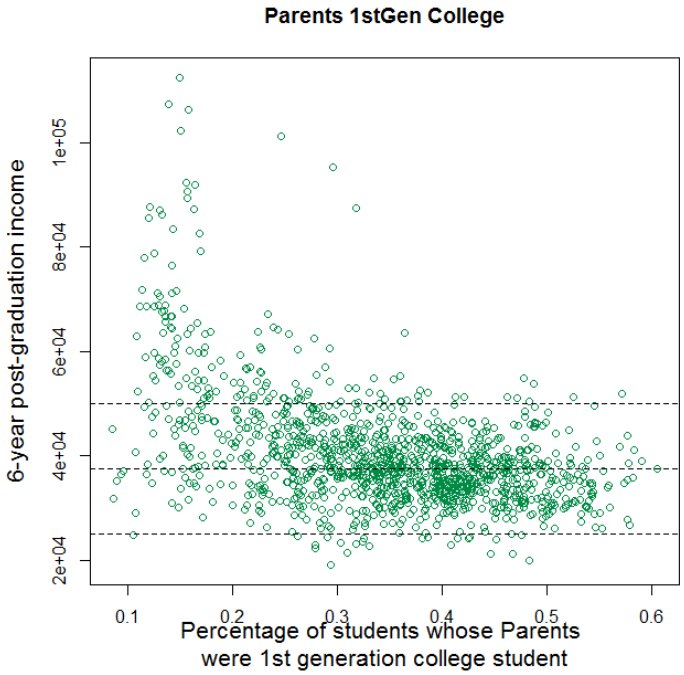

Figure 10. Relationship between six-year post-graduation income and percentage of students whose parents were 1st generation college students.

Thirdly, we found that the higher the proportion of female students in the cohort, the higher the chance that students are classified as Low income (see Figure 11). This reinforces prior findings that while there are now more women than men enrolled in higher education, a gender wage gap remains persistent in the U.S. labor market (DiPrete & Buchmann 2013; Kroeger et al. 2016). Part of the explanation may be a greater proportion of men enrolled in science, technology, engineering, and mathematic (STEM) disciplines that offer the highest salaries (Kim et al. 2015) and a concentration of men at the top of the wage distribution (Kroeger et al. 2016 p. 3). Besides, higher numbers of

Hispanic students in a cohort was predictive of more college graduates being classified as Very Low income, resonating with previous findings that young Black and Hispanic college graduates are more likely than their white counterparts to be under-employed (Abel & Deitz 2015; Kroeger et al. 2016). This could stem from concentrations of these students in particular types of institutions and fields of study, but may also be influenced by "discrimination or unequal access to the informal professional networks that often lead to job opportunities" (Kroeger et al. 2016 p. 12; Nunley et al. 2015).

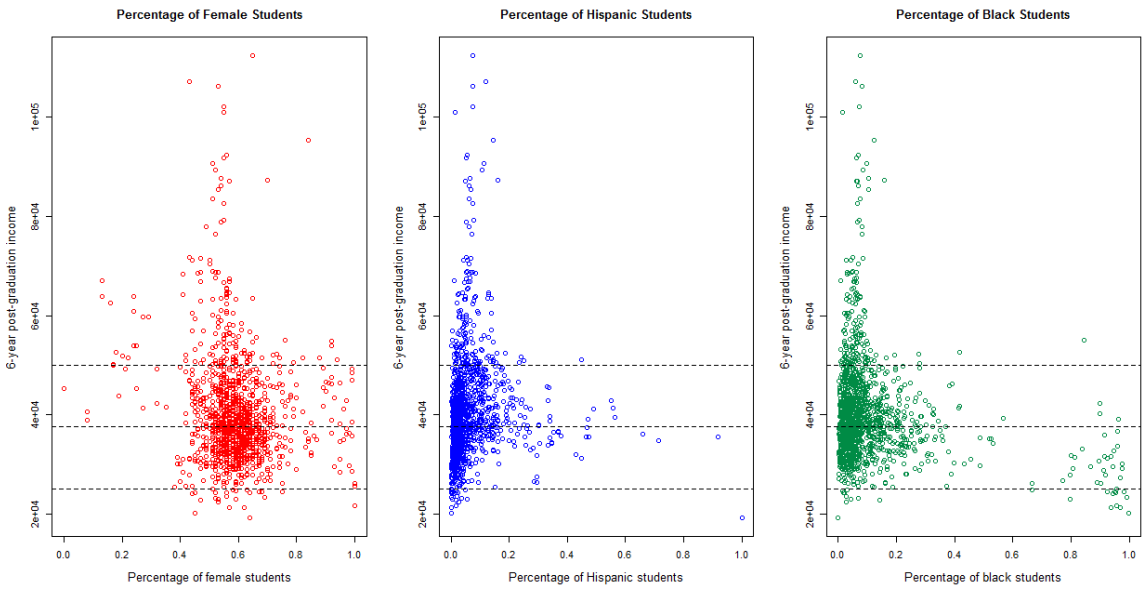

Figure 11. Relationships among six-year post-graduation income, percentage of female, Hispanic and Black students.

7. CONCLUSIONS AND FUTURE WORK

A college degree is perhaps the single best means for an individual in United States to enhance their income in the labor market. On the one hand, technological change and automation have tended to put downward pressure on the incomes of workers in low occupations (Brynjolfsson & McAfee, 2014). On the other hand, the globalization of production and offshoring has hit particular groups of non-degree holding "blue-collar" workers hard in recent years (Autor et al. 2016). It is clear, however, that a rising college wage premium has not meant that college graduates enjoy uniform access to highly paid work. Rather, the labor market for graduates is best viewed in terms of growing heterogeneity as the rewards have become unevenly distributed across the graduate population. This point is especially significant given the steadily rising costs of college over the past decades and associated increases in student debt (Avery & Turner, 2012). In this environment, the large scale of the College Scorecard dataset provides a highly important service by enabling prospective college students (alongside their parents and school advisors) to access transparent and detailed information about labor market outcomes of prior cohorts. We hope that our research complements the dataset by offering more insights of post-graduation incomes than the descriptive statistics provided on the College Scorecard website.

REFERENCES

Abel, J., & Deitz, R. (2015). Underemployment in the early careers of college graduates following the Great Recession. In *Education, Skills, and Technical Change: Implications for Future US GDP Growth*. University of Chicago Press.

Andrews, R. J., Li, J., & Lovenheim, M. F. (2016). Quantile treatment effects of college quality on earnings. *Journal of Human Resources*, *51*(1), 200-238.

Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175-185.

Altonji, J. G., Arcidiacono, P., & Maurel, A. (2015). *The analysis of field choice in college and graduate school: Determinants and wage effects* (No. w21655). National Bureau of Economic Research.

Autor, D. H. (2014). Skills, Education, and the Rise of Earnings Inequality Among the 'Other 99 Percent. *Science,* 344, 6186, 843-851.

Autor, D.H., Dorn, D., & Hanson, G. (2016). DP11054 The China Shock: Learning from Labor Market Adjustment to Large Changes in Trade.

Avery, C., & Turner, S. (2012). Student loans: Do college students borrow too much—or not enough?. *The Journal of Economic Perspectives*, *26*(1), 165-192.

Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2(2), 18-21.

Bartik, T. & Hershbein, B. (2016, June). *Degrees of Poverty: How Growing up Poor Changes the Returns to Education*. Paper presented at Association for Public Policy Analysis and Management International Conference. London School of Economics, UK.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. National Bureau of Economic Research.

Becker, G. S. & Tomes, N. (1994). Human capital and the rise and fall of families. In *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition)* (pp. 257-298). The University of Chicago Press.

Beaudry, P., Green, D. A., & Sand, B. M. (2014). The declining fortunes of the young since 2000. *The American Economic Review*, *104*(5), 381-386.

Beasley, J. E., & Chu, P. C. (1996). A genetic algorithm for the set covering problem. European Journal of Operational Research, 94(2), 392-404.

Blau, P. M. & Duncan, O. D. (1967). *The American occupational structure*. New York, London, Sidney: Wiley.

Blinder, A. S., & Krueger, A. B. (2009). *Alternative measures of offshorability: a survey approach* (No. w15287). National Bureau of Economic Research.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, 26(3), 801-849.

Brown, M., Haughwout, A., Lee, D., Scally, J., & Van Der Klaauw, W. (2014). *Measuring student debt and its performance*. Federal Reserve Bank of New York, Staff Report No. 668.

Brynjolfsson, E. & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company.

Cellini, S. R., & Chaudhary, L. (2014). The labor market returns to a for-profit college education. *Economics of Education Review, 43*, 125-140.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology,* S95-S120.

Cybenko, G. (1992). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems (MCSS), 5(4), 455-455.

Dale, S., & Krueger, A. B. (2011). *Estimating the return to college selectivity over the career using administrative earnings data* (No. w17159). National Bureau of Economic Research.

Deming, D. J., Yuchtman, N., Abulafi, A., Goldin, C., & Katz, L. F. (2016). The value of postsecondary credentials in the labor market: An experimental study. *The American Economic Review, 106*(3), 778-806.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2), 139-157.

DiPrete, T. A., & Buchmann, C. (2013). *The rise of women: The growing gender gap in education and what it means for American schools*. New York: Russell Sage Foundation.

Eide, E. R., Hilmer, M. J., & Showalter, M. H. (2016). Is it Where You Go or What You Study? The Relative Influence of College Selectivity and College Major on Earnings. *Contemporary Economic Policy, 34*(1), 37-46.

Federal Reserve Bank of New York. (2016). *The Labor Market for Recent College Graduates*. Retrieved online from https://www.newyorkfed.org/research/college-labor-market/index.html

French, M. T., Homer, J. F., Popovici, I., & Robins, P. K. (2015). What you do in high school matters: High school GPA, educational attainment, and labor market earnings as a young adult. *Eastern Economic Journal*, *41*(3), 370-386.

Golberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. *Addison-Wesley*.

Goldin, C, & Katz, L. F. (2009). The race between education and technology. Cambridge: Harvard University Press.

Gu, P., Zhu, Q., & Zhang, C. (2009). A multi-view approach to semi-supervised document classification with incremental Naive Bayes. *Computers & Mathematics with Applications*, 57(6), 1030-1036.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., et al. (2009). The WEKA data mining software. ACM SIGKDD Explorations Newsletter 11: 10.

Haughwout, A., Lee, D., Scally. J., & Van der Klaauw, W. (2015). *Press Briefing on Student Loan Borrowing and Repayment Trends, 2015*. Liberty Street Economics. Retrieved online from http://www.montana.edu/cstoddard/documents/Loans_and_Performance.pdf

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The Review of Economics and Statistics*, *91*(4), 717-724.

Holmes G., Pfahringer B., Kirkby R., Frank E., Hall M. (2002) Multiclass Alternating Decision Trees. In: *Elomaa T., Mannila H., Toivonen H. (eds) Machine Learning: ECML 2002*. ECML 2002. Lecture Notes in Computer Science, vol 2430. Springer, Berlin, Heidelberg.

Houle, J. N. (2013). Disparities in Debt Parents' Socioeconomic Resources and Young Adult Student Loan Debt. *Sociology of Education*.

Hout, M. (2012). Social and economic returns to college education in the United States. *Annual Review of Sociology*, *38*, 379-400.

Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2), 415-425.

Jones, J., & Schmitt, J. (2014). *A college degree is no guarantee*. Center for Economic Policy and Research. Washington, DC.

Kim, C., Tamborini, C. R., & Sakamoto, A. (2015). Field of study in college and lifetime earnings in the United States. *Sociology of Education*, *88*(4), 320-339.

Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. *In AAAI* (Vol. 2, pp. 129-134).

Kroeger, T., Cooke, T., & Gould, E. (2016). *The Class of 2016*. Economic Policy Institute. Washington DC.

Marginson, S. (2016). High participation systems of higher education. *The Journal of Higher Education*, *87*(2), 243-271.

Mettler, S. (2014). *Degrees of inequality: How the politics of higher education sabotaged the American dream*. Basic Books.

National Center for Educational Statistics [NCES]. (2016a). *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 2014*. Retrieved online from http://nces.ed.gov/programs/digest/d15/tables/dt15_302.60.asp?current=yes

National Center for Educational Statistics [NCES]. (2016b). *Percentage of recent high school completers enrolled in 2- and 4-year colleges, by race/ethnicity: 1960 through 2012*. Retrieved online from http://nces.ed.gov/programs/digest/d13/tables/dt13_302.20.asp

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2015). Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment. *The BE Journal of Economic Analysis & Policy*, 15(3), 1093-1125.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *The Journal of Higher Education*, 75(3), 249-284.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of democracy*, *6*(1), 65-78.

Putnam, R. D. (2016). *Our kids: The American dream in crisis*. New York: Simon and Schuster.

Rajput, A., Aharwal, R. P., Dubey, M., Saxena, S. P., & Raghuvanshi, M. (2011). J48 and JRIP rules for e-governance data. *International Journal of Computer Science and Security*, 5(2), 201-207.

Schofer, E. & Meyer, J. W. (2005). The Worldwide Expansion of Higher Education in the Twentieth Century. *American Sociological Review, 70*(6), 898-920. Psychology Press.

Thornton, R. J., & McDonald, J. A. (2015). The Gender Gap in Starting Salaries for New College Graduates. *Gender in the Labor Market (Research in Labor Economics, Volume 42) Emerald Group Publishing Limited*, *42*, 205-229.

U.S. Census Bureau. (2010). Distribution of Personal Income 2010. Retrieved online from http://www.census.gov/hhes/www/cpstables/032011/perinc/new01_001.htm

Vedder, R., Denhart, C., & Robe, J. (2013). Why Are Recent College Graduates Underemployed? University Enrollments and Labor-Market Realities. *Center for College Affordability and Productivity (NJ1)*.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wolff, E. N., Baumol, W. J., & Saini, A. N. (2014). A comparative analysis of education costs and outcomes: The United States vs. other OECD countries. *Economics of Education Review*, *39*, 1-21.

APPENDICES

Appendix A. Potential Attributes.

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
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