BridgeEdU Challenge: Emergency Financing Eligibility and Supports via Student Persistence Scoring

The authors of this paper work in the machine learning and data analytics industry. Previous research by the authors include image recognition, fraud detection, and healthcare data analysis using machine learning. The authors are willing to continue to explore the data for BridgeEdU after the competition and are open to providing consulting services.

Executive Summary

Student persistence is a metric scrutinized by both the government and academic communities interested in improving education. Major works to understand persistence by the Obama Administration and other scholars have so far been incomplete. In fact, these reports do not take into consideration important factors such as first-generation college students, financial aid packages, and factors outside of the immediate classroom setting (e.g. having children). Even the studies by Princeton's National Longitudinal Study of Freshmen use outdated logistic regression methods that yield low prediction power. The following study combines the available dataset and used socioeconomic frameworks created by previous research to determine high-risk population most prone to dropout. Using a machine learning algorithm called random forest, the model was able to predict 73.87% of students who are likely to drop out. This is a significant improvement to the existing logistic regression methods that yielded a poor 18.01% prediction accuracy. The following report summarizes the 21 most important factors including GPA, perception of prejudice on campus, and percentage of classes dropped and ranks them to create a potential student persistence score. A proposed plan of action is attached to utilize the random forest algorithm for BridgeEdU to identify qualifying students for the emergency gap fund.

Introduction

Student Persistence Score (SPS), defined as a student's likelihood of completing a postsecondary degree, is a metric of high interest but with little quantifiable predictors. While college persistence is increasingly discussed more at a national level, its indicators and accompanying policies are greatly misunderstood and at times detrimental. For example, proposals to reward or punish institutions for their graduation rate capture the misuse of data about persistence and further hinder students' success in college.[1] The following report examines previous work on student persistence both by the government and private institutions to statistically determine pertinent factors contributing to persistence. The result is presented as a potential algorithm to identify students who qualify and will benefit the most from emergency gap financing.

Previous Studies

During the Obama Administration, the federal government has pushed to document and understand student persistence across racial, ethnic, and gender groups. *Higher Education: Gaps in Access and Persistence Study* by the National Center for Education Statistics (NCES) identified 46 indicators of persistence and recommended changes for the Higher Education Opportunity Act.[2] The findings in the NCES report, however, are limited as its focus was only to explore the variables leading to the persistence gap in sex, racial, and ethnic differences. It simply summarizes the statistical differences and fails to provide the variables affecting these results. Another work by the White House in 2014 examined the early indicators of college readiness, and in conjunction, persistence.[3] It found that student achievement in 8th grade and enrollment in remedial courses as the two main factors impacting college readiness, which it projected to persistence levels.[4,5] However, the correlation between college readiness and persistence was not established quantitatively or discussed thoroughly, so further investigation was needed.

Apart from the government, the AFT Higher Education Council in 2003 analyzed the six-year longitudinal study of college students by NCES from 1995 to 2001.[6] This report identified several factors including part-time enrollment, delaying entry into college, not having a regular high school diploma, having children, being a single parent, being financially independent of parents, and working full-time while enrolled as significant indicators of student persistence. While these risk factors are important, this study does not differentiate between grade levels, effects of being a first-generation student, financial aid, and race in a meaningful way.

Lastly, comparison studies on the determinants of persistence for first-generation and continuing-generation students are well-documented by Mandy Martin Lohfink and Michael B. Paulsen.[7] This study summarizes the negative effects race, parental education status, and gender have on persistence levels for first-generation

students. It also excellently sums up the effects of academic performance, educational aspirations, work-study aid availability, and satisfaction with social life as main indicators for persistence. Interestingly though, they found that none of the precollege achievement variables were significantly related to first-to-second-year persistence—perhaps contradicting what was found in the White House report. These results imply that none of these frameworks and previous works capture a complete overview of student persistence.

Theoretical Framework

The following exploration of the data builds on the examination of student persistence mentioned above. While our approach was mainly driven by the principles of statistical modeling and machine learning techniques, we included the sociology and economics theoretical frameworks included in previous works. Namely, sociological theory from the past 40 years (Alexander & Eckland, 1975; Blau & Duncan, 1967; Eckland & Alexander, 1980; Inoue, 1999; Kalmijn, 1994; Lampard, 1985; Parsons, 1959; Sewell & Shah, 1967; Sewell & Hauser, 1975; Thomas, Alexander, & Eckland, 421 NASPA Journal, Vol. 41, no. 3, Spring 2004 1979; Trent & Medskar, 1968; Updegraff, 1996; Wolfle, 1985) and economic theory related to education and financial aid (Becker, 1993; Denison, 1964; McPherson, 1982; Schultz, 1995; Corrazini, Dugan, & Grabowski, 1972; Hoenack & Weiler, 1975).[8] These theories guided the research to consider the effects of background, family, academic ability, and aspirations for social impact on persistence, as well as financial aid and financial independence for economic links to SPS.

However, these theoretical frameworks were only applied after looking at the statistical measures of the available dataset. The research was guided by supervised learning algorithms to remove bias from the dataset, and corresponding patterns gleaned from the data were then explained by said frameworks. This meant evaluating multiple data sources for their size, breadth of field, and term length before drawing conclusions. Data sources needed to be large enough to train the algorithm that utilized machine learning techniques. The breadth of data collected in each source needed to be broad enough to cover a variety of relevant topics to student dropouts. This included financial, educational, and background data. Additionally, the data needed to be screened such that any fields used in the model represented data that could be feasibly obtained by BridgeEdU. Finally, the chosen data set needed to follow students starting before they entered college until they either graduated, transferred, or dropped out.

Research Questions

The following research questions are addressed in this study.

1. What factors impact SPS from a statistical modeling standpoint?
2. How accurate are these predictive models and how do they compare to previous work?
3. In applying the sociology and economics frameworks, are these results plausible and can it be used by BridgeEdU to apply SPS for emergency gap financing?

Methodology

This study extensively used the dataset collected by the Princeton Office of Population Research.[9] Building on the works of Cox et al. on this dataset, the study first determined the important factors laid out by previous research and applied a different machine learning technique called random forest to validate the findings.[10]

Data

The dataset examined in this research was from the National Longitudinal Study of Freshman collected by Princeton between 1999 and 2003. This specific set was used, because it was more recent than the NCES data and had more variables to explore than others previously mentioned. The data was collected in five rounds from a total of 3,924 students. The first round was collected at the beginning of freshman year, and subsequent rounds were collected during the spring semester of each year of college.

The data collected spanned several different verticals:

- The first vertical was academic performance. This covered overall GPA, grades in specific classes, time spent studying, and time spent working on homework.
- The second vertical was campus life. This covered time spent with friends, how lonely the subject felt, whether or not the subject had been the subject of prejudice, and ethnic breakdown of the campus.
- The third vertical was financial status. This included family income, family contribution, subject income, financial aid (loans and grants), and family homeownership.
- The fourth vertical was the subject's background. This included information from all the other verticals when the subject was 6-10 years old, 13 years old, and during their last year of high school. In addition, this data covered information about the subject's neighborhood, home life, and parental involvement in the subject's schooling, all at the previously mentioned ages.

Preprocessing

To prevent against overfitting and to focus the model on previously determined socioeconomic factors, the dataset was first compared against the findings of Cox et al. This study found that financial factors did not have a significant effect on persistence while gender, race, academic performance, attitudes toward education and the student's college, and psychological stresses did. These findings are consistent with the works done on the NCES study.

Next, dropout indicator was defined to be students who did not have a graduation record or those who had not yet attained a degree in 6 years to model SPS against a certain benchmark. This metric was also used in both the Cox et al. and the NCES study. The data suggests that most dropouts occur in the last two years of college. In total, only 42 students (1.1%) in the data had dropped out before the end of sophomore year, whereas 464 students (11.8%) had dropped out junior year and beyond. Then, all the attributes for students that had not dropped out in the first two years were aggregated to predict whether or not the respective student would drop out in their following two years and beyond.

In examining the data further, it was determined that the median household income for students was $50,000 - $75,000 in 1999, when the nationwide median household income was $41,994. Thus, the students in this dataset were skewed towards those from higher income backgrounds, which may explain why previous studies did not find significant financial effects. To build a predictive model that effectively includes financial information as a relevant contributing input, the dataset was filtered to adjust for this factor, which was not included in previous works that focus primarily on college costs and household income. From the surveys, the total cost of college for each student (tuition, room and board, etc) and the total monetary contributions to cover those costs (parents, loans, grants, etc) were calculated. Those values determined the funding gap between total costs and contributions for freshmen and sophomore years. This revealed a subset of students who would have experienced a potentially high level of financial stress due to $20,000 or more in funding gap. Amongst these 465 students, nearly a quarter of them eventually dropped out, indicating that this subset was the group of interest.

Statistical Method

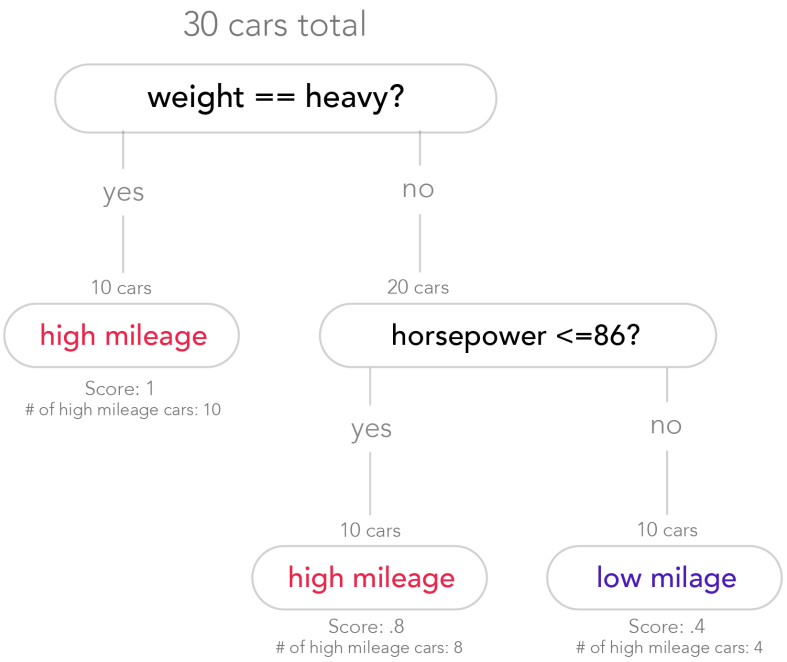
Previous work on student persistence have primarily used logistic models to predict the likelihood of dropping out

or to explain the effects of different variables on persistence. While logistic regression fares relatively well in simple supervised learning models like the dataset of interest, the specific models reported did not yield strong predictive power. The two logistic models in the Cox et al. paper, which used the same dataset as this study, had R^2 values ranging only from 0.187 - 0.240. Similarly, Somers et al. paper that studied the NCES data used logistic regression and yielded R^2 values between 0.1096 - 0.1206. R^2 is one measure of how well a statistical model explains the data. Its values range from 0 to 1 indicating how much of the data the model is able to explain with 0 being the least and 1 being the most.

To improve upon models that aim to predict student persistence, this study utilized a machine learning algorithm called random forest. Random forests have been used extensively in both industry and research settings with a high degree of accuracy compared to other techniques. For example, in the healthcare research field--which traditionally yielded poor predictive performance--random forests have been used to diagnose Alzhemier's disease based on MRI data with 90% accuracy and have predicted patient outcomes to clinical drugs with more than 83% accuracy.[11]

Decision Trees

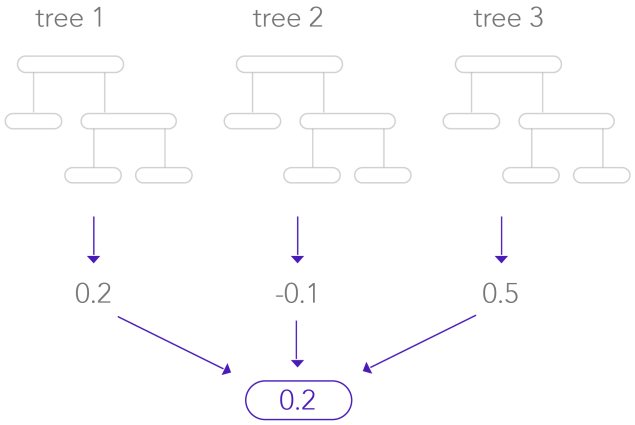
To understand random forests, one must first understand decision trees. To illustrate the underlying algorithm behind decision trees, take the following example: is a car a high mileage (low mpg) vehicle? (Score of 1 = yes, 0 = no).



A decision tree is a series of yes or no rules that splits the data points into different classification buckets. When a new observation is made, it is then passed down the tree and iterates through the series of binary rules (yes or no) until it reaches the bottom of the tree where each classification bucket cannot be broken down further. In the example above, a decision tree has been constructed based on the data of 30 observed cars. Given a new unknown car (31st data point), the algorithm can determine whether or not it is low mileage (high mpg) or high mileage (low mpg) based on the given attributes on the branches. For example, if the 31st car is heavy, then the algorithm would predict it as high mileage with a probability of 1.0. If the car was not heavy and had horsepower greater than 86, the algorithm would predict it as low mileage with probability of 0.4.

Random Forest

A random forest generates multiple decision trees, oftentimes hundreds or thousands, by randomly selecting a subset of variables and observations to construct each individual tree so that each tree will also have different yes or no rules at each break. Then, when a new prediction is needed for a new observation, that data point is passed through each of those individual decision trees, like in the example above, and the scores are averaged to obtain a final prediction.



An intuitive way to understand why random forest achieves better results than a single decision tree is to imagine each decision tree as a single voter. A single voter is less likely to make a correct prediction than hundreds of voters who each have different and diverse perspectives of the data set. Because each decision tree in the random forest has not seen all of the data and variables, random forests are also less likely to overfit the data than a single tree which has been perfectly tuned to a training data set.

Results

Given the target population, the algorithm iterated through different combinations of factors to determine the statistically most predictive model. Random forest indicated the following 21 factors were the most predictive in the order of importance measured by its effects on the accuracy of the model:

Variable	Description	Variable Importance
GPA		0.145
Perception of Prejudice On Campus	Survey Question: rank 0-5 how often you experience various acts of prejudice during the school year. We aggregated all of these score over a 2 year period.	0.060
Percentage of Classes Dropped		0.059

Variable	Description	Variable Importance
Percentage of Students at University Who Applied for Financial Aid		0.058
Classes Completed		0.052
Total Contributions From Parents	Dollar amount contributed by parents toward college expenses	0.051
Average Hours Per Week Studying		0.050
Average Hours Per Week Spent On Extracurriculars		0.049
Level of Effort Placed Into Studies		0.048
Total Cost Gap	Cost of college minus total funding sources	0.044
Sentiment Toward Spending Time Away From Campus	Survey Question: I just wanted to get away from campus for a while. Rank how often you felt this, 0-10	0.043
Household Size		0.042
Total Funding Gap Percent	Percent of total college expenses	0.042
Total College Cost		0.037
Household Annual Income		0.036
Total Grants Received		0.036
AP Tests Passed		0.031
Hours Per Week Working Freshman Year (Job)		0.031
Loneliness/Homesickness	Survey Question: I felt lonely and homesick. Rank how often you felt this, 0-10	0.030

Variable	Description	Variable Importance
Total Contribution	Total contributions toward college expenses	0.030
Hours Per Week Working		0.030
Sophomore Year (Job)		0.030

The corresponding model was cross validated by randomly splitting the dataset of 464 students with funding gaps greater than $20,000 into two 70/30 groups. The group with 70% of the observations was used to build the model, also known as a training set to train the model. The remaining 30% tested the predictive power of the model as these data points have not been seen by the model. The model gives a score ranging from 0.0-1.0 of how likely a particular student will drop out with 0.0 being least likely and 1.0 being most likely.

The results of the random forest model compared to a logistic regression model using the same variables is summarized below:

	Random Forest	Logistic Model
Dropouts Correctly Predicted	**73.87%** **(82 out of 111 dropouts)**	**18.01%** **(20 out of 111 dropouts)**
Overall Accuracy	72.69% (338 out of 465 predicted)	76.13% (364 out of 465 predicted)
R Squared Value (higher better)	0.881	0.761

Discussion

From the table above, random forest and logistic models perform very similarly in terms of overall accuracy. At a first glance, it might seem that the logistic model has the edge in terms of accuracy; however, that is primarily due to the fact that roughly three quarters of the students did not dropout. If one were to randomly guess which students dropped out, they would be right roughly 75% of the time, thus the 76.13% overall accuracy for the logistic model is not outstanding from that perspective. What is more important and impressive is how many dropouts the random forest model is able to correctly predict since that set of students is in the minority of the overall data set and as a result is harder to identify.

Thus the key metric to focus on is how many dropouts each model was able to correctly predict. Out of the 111 students who dropped out, the logistic model was only able to predict 20 of them while the random forest was able to correctly identify 82. The results above demonstrate that the random forest model is a clear improvement over

the previous models used to predict student persistence given that it is able to predict a significantly larger number of students who drop out. Such a model will help BridgeEdU better identify students who are at risk of dropping out.

For the students the model correctly predicted to drop out, all of them had a random forest score greater than 0.26 on a scale of 0 - 1.0. Thus a useful implementation would be to score each student using this model, identify those with scores near or at 0.26, and select those specific students for further examination as potential candidates for receiving the BridgeEdU bridge loan or grant. Then those students can be passed through the model again with the additional loan/grant money to see if that would bring their score below the 0.26 threshold; if it does, then that particular student would be a strong candidate for the loan/grant.

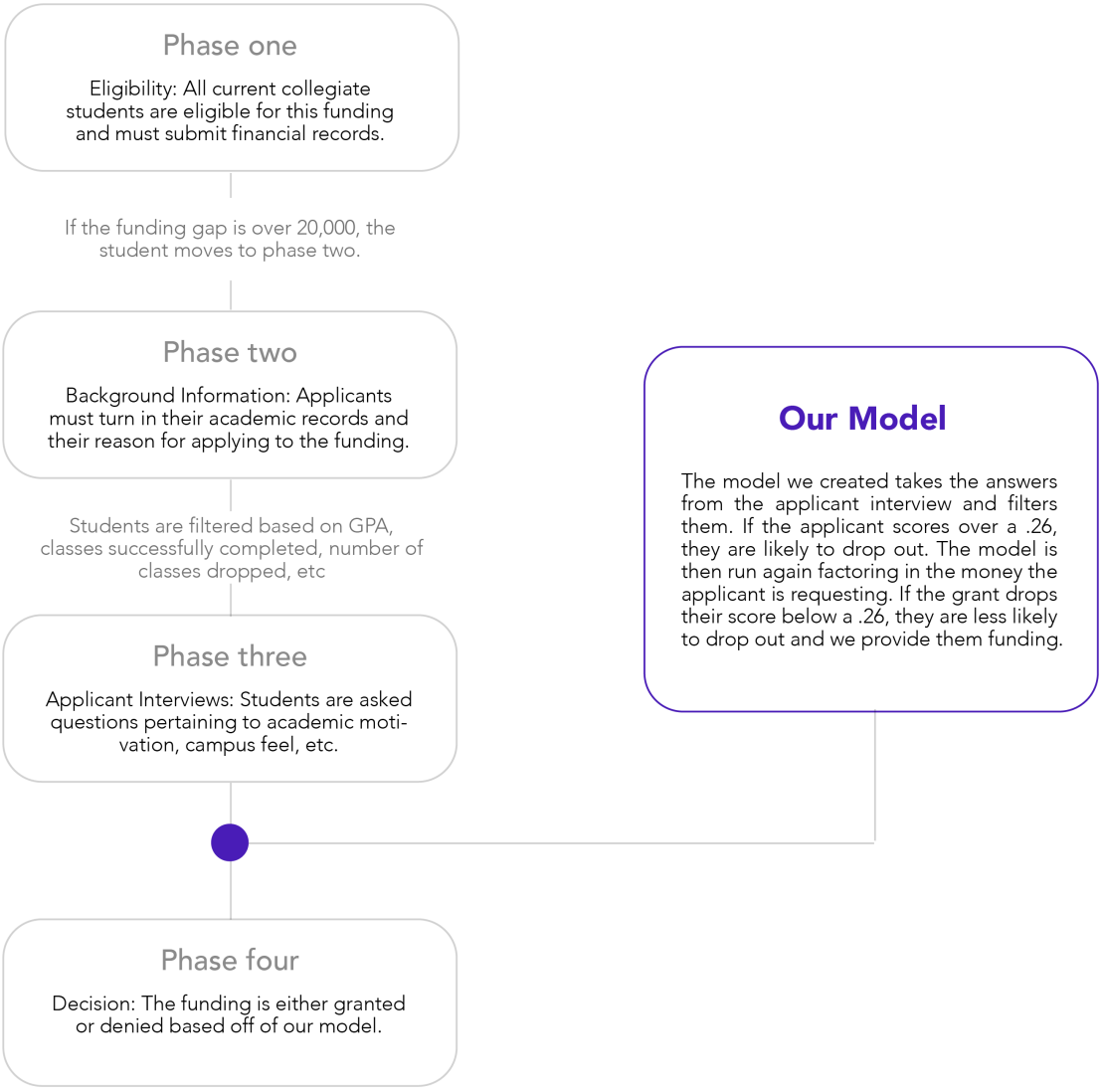
A limitation of this model is that the dataset used was comprised entirely of self-reported answers from a survey. Respondents' ability to recall certain details and interviewer bias may have impacted some of the responses. Furthermore, the data set was relatively small, just 465 students who met the criteria of having $20,000 or more in funding gap. The data was also recorded from 1999-2003, more than a decade and a half ago. These factors question the validity and applicability of the dataset in current times as the national dropout rate was 55% in 2014 according to The National Student Clearinghouse while our dataset only had 12.9% of students dropout. A more balanced dataset would improve modelling and produce more accurate results. Future steps would be to collect as much data from recorded sources such as transcripts and financial documents as possible, to increase the data set size, collect on a more representative sample of college students, and collect more recent data.

Proposed Solution

The proposed means of collecting the information necessary to drive the solution described above is two-fold. Data will be collected either through BridgeEdU's pre-existing first-year experience platform or through the application process. In either case, the data will be collected from the same sources. Academic performance data such as GPA, percent of classes dropped, number of completed classes, and number of AP classes completed in high school can be gathered from a student's transcript. All of the financial data including the funding gap, total parent contribution, total college funding, total dollar amount of grants received, and total college expenses, can be gathered from a combination of a student's FAFSA and billing statements. Finally, all of the subjective and self-reported data such as perception of prejudice on campus, hours/week spent on extra-curriculars, and amount of effort put into class will be collected in person in an interview setting or online through surveys.

In order to make the process of collecting student data and selecting recipients of funding as efficient and time-effective as possible, it is proposed that the data be collected in three stages. The first stage would consist of gaining access to student's financial information through partnering with the FAFSA Office. As a potential provider of student aid, it is possible to have students' financial information disclosed directly. As the model used in this proposal suggested that a funding gap of greater than $20,000 was an indicator of lowered retention, students fitting this criteria could be targeted for further data collection. The second stage would consist of an online application process during which students would disclose further financial details, such as their immediate need for gap financing, and academic information, i.e. their transcript. The final stage would consist of an in-person interview where all of the self-reported data would be collected.

This progression would allow students to be filtered in three places. Students would be first filtered based on their probable need for financing as defined by the size of their funding gap. They would then be filtered after the online application process based on the model described in the prior sections, in addition to responses with respect to their need for gap financing. Finally, students would be filtered after an in-person interview during which more subjective data, i.e. level of motivation, work-life balance, and extenuating circumstances, would be collected. This multi-step process would ensure that all easily collected, valuable data is gathered as early as possible, ultimately reducing the effort necessary to select candidates for the gap financing. The data from the first two steps is easy to collect and aggregate, while only the third step of data collection will take a significant amount of time and effort. This should be a reasonable expectation of due diligence when screening students for this program.



Conclusion & Further Discussion

SPS is a hard metric to algorithmically define. While the Princeton data and the NCES data discussed here provide excellent indicators, it is inherently limited by the quality of self-reported scores. Still, the study provided here was able to improve the prediction percentage by 55% and give a more modern approach to determining SPS values that is consistent with the techniques used in industry and research today. Provided more time and resources, the authors want to collect more recent data to train the algorithm with current trends. Also, 465 points used in this study is relatively small. Having more data can help improve the accuracy of the algorithm. Lastly, if given more time, deep learning methods that find patterns in student persistence data may be leveraged to overcome the bias that is inherent with tagging survey data. Deep learning methods are at the forefront of artificial intelligence and machine learning research, and it may benefit algorithm development with SPS as well.

Footnotes

1. Gold, Lawrence N. et al. “Student Persistence in College: More Than Counting Caps and Gowns” AFT Higher Education Program and Policy Council. Aug 2003.

2. Ross, Terris. Kena, Grace. Rathbun, Amy. KewalRamani, Angelina. Zhang, Jijun. Kristapovich, Paul. Manning, Eileen. “Higher Education: Gaps in Access and Persistence Study. Statistical Analysis Report” National Center for Education Statistics and American Institutes for Research. Aug 2012.

3. “Increasing College Opportunity for Low-Income Students: Promising Models and a Call to Action” The Executive Office of the President. Jan 2014.

4. ACT, “The Forgotten Middle: Ensuring that All Students Are on Target for College and Career Readiness before High School,” 2008.

5. Complete College America. Report from 31 Complete College America Partner States. 2014.

6. Gold, Lawrence N. et al. “Student Persistence in College: More Than Counting Caps and Gowns” AFT Higher Education Program and Policy Council. Aug 2003.

7. Lohfink, Mandy Martin. Paulsen, Michael B. “Comparing the Determinants of Persistence for First-Generation and Continuing-Generation Students.” Journal of College Student Development. July/Aug 2005. Vol 46, No 4.

8. Somers, Patricia. Woodhouse, Shawn. Cofer, Jim. “Pushing the Boulder Uphill: The Persistence of First-Generation College Students.” NASPA Journal, Vol. 41, no. 3, Spring 2004.

9. Massey, Douglas S. Charles, Camille Z. National Longitudinal Survey of Freshmen. Office of Population Research. Princeton University. 2008.

10. Cox, Bradley E. Reason, Robert D. Nix, Samantha, Gillman, Megan. “Life Happens (Outside of College): Non-College Life Events and Students’ Likelihood of Graduation” Research in Higher Education. Nov 2016, Vol 57, Issue 7, pp 823-844.

11. Lebedev, Alexander. Westman, Eric. Wesen, G. J. P. Van. Kramberger, M. G., Lundervold, Arvid. Aarsland, Dag. Soininen, H. Kloszewska, I. Mecocci, P. Tsolaki, M. Vellas, B. Lovestone, S. Simmons, Andrew. “Random Forest ensembles for detection and prediction of Alzheimer’s disease with a good between-cohort robustness” NeuroImage: Clinical 6 p. 115-125. 2014.