Census Data, Administrative Records, and Machine Learning: Exploring Micro and Macro Impacts on Four-Year Graduation Rates

Mathew E. Hauer
Public Service Associate
Carl Vinson Institute of Government
University of Georgia
201 N Milledge Avenue
Athens, GA 30602
hauer@uga.edu | 706.542.9369

James M Byars
Scientific Computing Professional Associate
Carl Vinson Institute of Government
University of Georgia
201 N Milledge Avenue
Athens, GA 30602
jmbyars@uga.edu | 706.542.5781

Keywords
College access, College success, Retention, Graduation, Progression, Machine Learning, Random Forest

Abstract

College accessibility is near an all-time high, in spite of this improved college access, four-year college cohort graduation rates are still below 40% in the United States. Theoretical models of student persistence – the general progress toward obtaining a degree – find that a combination of individual characteristics and educational institution factors drive four-year graduation rates. However, the effect of a student’s community of origin at the time of enrollment on four-year graduation rates has been unexplored. Here we combine complete administrative student enrollment records for the thirty-three member postsecondary education institutions in the University System of Georgia (n = 110,576) with detailed demographic information from the US Census Bureau to merge administrative student-level data with macro US Census information. We use a random forest model to predict the four-year graduation rates of the 2008 - 2010 first-time freshmen cohorts. Our preliminary results suggest accurate prediction of whether a student graduates 83% of the time.
Introduction

Despite improved college access, a large number of students still fail to complete degrees. Of the 1998 college cohort, for instance, only 35 percent obtained a bachelor degree in four years, and only 57 percent in six (Knapp, Kelly-Reid & Whitemore, 2006). To make matters more complicated, the widening degree completion gap between high socioeconomic status (SES) students and low SES students presents another challenge to achieve state postsecondary education goals (Titus, 2006; Walpole, 2003). Aside from SES status, differences in degree attainment along racial, ethnic, and gender lines, highlight additional sources of educational inequality (Fischer, 2007; Gloria, Robinson Kurpius, Hamilton, & Willson, 1999; Leppel, 2002; Muñoz & Maldonado (2012).

The complexity of college student retention, progression, and graduation (RPG) has become a popular research topic among institutional researchers, postsecondary education scholars, and policy makers in recent years. National initiatives such as Lumina’s Goal 2025 and Complete College America emphasize not only increased statewide degree attainment, but postsecondary education infrastructures and policies that encourage judicious time-to-degree patterns among college-going students. Currently, 33 states and the District of Columbia participate in the Complete College American initiative which encourages state political and higher education leaders to set statewide completion goals, expand data collection procedures on RPG outcomes, and take political action to improve the state’s overall degree attainment. Student retention and graduation have been the primary focus of scholars with relatively few scholars focusing on progression.

The current empirical research on student RPG has focused exclusively on student-level and institutional-level factors, drawing upon various disciplinary perspectives such as sociology, economics, and psychology. For example, Tinto’s (1975, 1993) seminal study on student integration has had a profound effect on the direction of theoretical and practical approaches to student persistence research since its original publication. Drawing upon Durkheim’s theory of suicide, Tinto assumes an interactional perspective, viewing college persistence as a dynamic and reciprocal process between the individual student and the college environment (1975, 1993). A student’s academic commitment, academic expectation, financial resources, and pre-college education interact with an institution’s social and academic environment. The integration between student and their college environment is characterized by academic integration and social integration. Academic integration occurs when students develop a significant attachment to academic purpose of attending postsecondary education; whereas, social integration occurs when students forms relationships and experiences beyond the classroom. According to Tinto, academic integration directly influences students’ goal commitment while social integration directly influences their commitment to the institution. Therefore, both are essential for student persistence under Tinto’s student integration model. However, a major limitation of Tinto’s theory is that it does not explicitly address the role of exogenous factors such as community, political, and economic dynamics that influence, shape, or disrupt a student’s integration into a postsecondary education institution.

This study seeks to address the gaps in the substantive and statistical considerations of student RPG outcomes. As referenced earlier, the current state of the student RPG literature isolates
student and institutional characteristics, but largely ignores the historical and environmental contexts which influence students beyond college campuses. More specifically, the influence of a student’s home community is largely ignored and serves as a valuable research opportunity when increasing our understanding of student access and success. Utilizing individual student record administrative data from thirty-three member institutions of the University System of Georgia combined with US Census data, this study addresses this literature gap by linking student enrollments records to quantitative measures on the communities from which students originate. We utilize a machine learning approach in a post-hoc analysis that focus on whether future students will graduate in a timely manner. Understanding a model’s predictive ability allows institutional researchers and administrators along with state educational and political leaders to address opportunities for further investment, policy and program re-evaluation, and continued support to achieve state degree attainment goals.

Data

This study analyzes state and federal administrative datasets in order to study micro- and macro-level factors impacting college student RPG. Micro-level factors include characteristics of students first enrolling at a public four-year postsecondary education within the state of Georgia from fall 2008 through fall 2012. This data was provided as a part of the collaboration with the University System of Georgia’s Board of Regents and includes variables about students’ postsecondary education enrollment and outcomes, demographic characteristics, previous academic credentials (e.g. secondary education institution attended), and financial aid disbursements. Macro-level effects are captured through the 2012 five-year estimates of the Census Bureau’s American Community Survey (ACS) for all geographies. Using this specific ACS release benefits this study’s analytical strategy in two ways. First, the 2012 ACS aligns with the most recent four-year undergraduate cohorts available from the USG dataset. Second, this release is centered on the 2010 US Census which improves survey data estimates because the sample weights are calibrated against the 2010 Census. The Census provides more stable and reliable estimates because of its comprehensive sampling strategy and increased number of respondents.

These two datasets were merged by linking a student’s geographic location with their associated 2010 Census block group designation. This was achieved by geocoding student mailing addresses that were listed on their first matriculation term enrollment record. These observations expressed in latitude and longitude coordinates were then spatially joined to a Census block group that it resides in using ESRI’s ArcMap. The block group is the smallest geographic unit at which the Census Bureau publishes sample data. The main benefit of including variables provided by the 2012 ACS is that we can capture many of the nuances associated with a student’s community upbringing such as community educational attainment, socioeconomic status (e.g. median income and poverty levels), and the work force composition. Past studies on student RPG have failed to capture this essential factor in the student access and success literature.

This dataset includes information on 110,576 enrolled undergraduate students and over 3,000 predictors. Preliminary results examined a redacted dataset to reduce computational demands.
Future model specifications will incorporate all available data points in the USG and ACS datasets and will be estimated using available high-performance computing hardware.

**Bibliography**


Figure 1. Model Results.

ROC (continuous) and ROCH (dotted)

H measure w(c)

AUC w(c)

Smoothened score distributions (class 0: dash-dotted, class 1: dashed)
Table 1. Variable Importance.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Did Not Graduate</th>
<th>Graduate</th>
<th>Mean Decrease Accuracy</th>
<th>Mean Decrease Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>High School GPA</td>
<td>247.44</td>
<td>73.11</td>
<td>209.14</td>
<td>4918.42</td>
</tr>
<tr>
<td>25+ with Bachelor's Degree or Higher</td>
<td>66.94</td>
<td>73.72</td>
<td>127.8</td>
<td>3090.54</td>
</tr>
<tr>
<td>SAT Score</td>
<td>104.21</td>
<td>25.24</td>
<td>126.6</td>
<td>3030.2</td>
</tr>
<tr>
<td>Median Household Income</td>
<td>52.37</td>
<td>81.29</td>
<td>113.56</td>
<td>2974.94</td>
</tr>
<tr>
<td>Distance to School</td>
<td>96.75</td>
<td>11</td>
<td>100.2</td>
<td>3226.87</td>
</tr>
<tr>
<td>School Sector</td>
<td>69.51</td>
<td>12.28</td>
<td>83.31</td>
<td>895.34</td>
</tr>
<tr>
<td>Race</td>
<td>47.3</td>
<td>43.31</td>
<td>77.68</td>
<td>785.7</td>
</tr>
<tr>
<td>Degree Code</td>
<td>45.3</td>
<td>-40.08</td>
<td>17.57</td>
<td>147.33</td>
</tr>
</tbody>
</table>

Table 2. Confusion Matrix of Validation Data Set.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted Graduated</th>
<th>Did not Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduated</td>
<td>1,562</td>
<td>2,258</td>
</tr>
<tr>
<td>Did not Graduate</td>
<td>792</td>
<td>11,746</td>
</tr>
</tbody>
</table>