SMU Data Science Review

Manuscript 1060

Postsecondary Attainment: Identifying Areas to Improve Retention for North Carolina Community Colleges

Noelle Brown

John Heinen

Matthew Rega

Lizzy Sterling

Jacob Drew

Postsecondary Attainment: Identifying Areas to Improve Retention for North Carolina Community Colleges

Noelle Brown, John Heinen, Matthew Rega, Lizzy Sterling, Jacob Drew

Master of Science in Data Science, Southern Methodist University,
6425 Boaz Lane, Dallas, TX 75275 USA
{noelleb, jheinen, mrega, lsterling, jdrew}@smu.edu

Abstract. This paper details ways that North Carolina Community Colleges can improve retention by investigating curriculum completion and success in students' first year of college. A complete repository of data from community colleges in North Carolina was built for this analysis featuring 1950 variables related to the colleges and their surrounding communities from five main publically available data sources. Leading factors were identified to help explain why one-third of students in North Carolina community colleges do not return for their second year of college [2]. The results revealed that racial demographics along with success from educationally at risk and English as a Second Language (ESL) individuals receiving Adult Education and Family Literacy Act (AEFLA) services are leading factors to predicting retention. Our framework included elastic net regularized regression techniques to determine the feature importance of variables connected to students' success in their first year of college.

1 Introduction

It is understood that postsecondary education is a useful tool to get ahead in today's workforce. In 2017, 83% of North Carolina public high school students intended to go to college [2]. This number is down from a peak of 86% in 2010 [2]. The number of public high school students that actually enroll in college is much lower. Only 62% of these students end up enrolling in college, according to numbers from 2014 [2]. Of these students that enrolled in the fall of 2014, only 77% ended up returning for their second year of school [2]. This all leads to less than half of North Carolina's adults with a postsecondary degree or certificate as of 2016 [2]. To achieve a statewide goal of 60% of North Carolina's population holding a postsecondary degree or professional certificate, 672,000 more working-age North Carolina residents need to complete a higher education degree or credential program [2]. At current rates it is believed that the postsecondary attainment gap would shrink to 420,000 by 2025 [2]. This is not a simple task due to the variability in postsecondary attainment that exists between residents of various demographic and socioeconomic statuses. Out of 100 counties in the state, only six meet or exceed the statewide goal of 60% [2]. Less than 47% of adults hold a postsecondary degree or high-quality credential in the remaining

counties [2]. The gap between counties is substantial with at least five counties below 20% attainment [2]. This analysis uses data related to community colleges in North Carolina to identify factors that could help increase postsecondary educational attainment. By ensuring that students are successful in the first year of college, North Carolina can work toward reaching this attainment goal.

First year retention rates in North Carolina in 2015 stood at 77% [2]. A closer look at this number reveals that while first year retention rates for schools within the University of North Carolina system have retention rates at 87%, North Carolina Community Colleges are at 65.7% [2]. This is a downward trend from 68% in 2007 [2]. Additional research from the National Center for Education Statistics does not point to a single reason for why students do not return for their second year [2]. In the 2003-04 school year, students at all types of colleges could choose multiple reasons for why they did not return for their second year: "Personal reasons were the most commonly cited (53%), followed by financial reasons (31%), family responsibilities (21%), lack of satisfaction (17%), and academic problems (13%)" [2].

The first step in the process of identifying factors leading to successful first year retention in North Carolina Community Colleges was to compile a repository of data that could be relevant to the areas of opportunity and current success at these schools. Five categories of data were identified and gathered. These five categories included:

- Population, graduation, and first year success
- Assistance Services for students needing help to succeed
- Surrounding Community metrics
- Community Individual/Family Income metrics
- North Carolina Public High School metrics

Combining these data sources provided a rich data set to conduct research from different perspectives. Thousands of variables were made available through the use of these sources.

Once the data repository was created, the next step was to identify a variable of interest. This variable was first year progression rate. First year progression rate is defined as the percentage of students that attempt and complete at least 12 hours of classes within their first academic calendar year. For those students that do enter community college, first year progression is the first step towards success, which includes staying in school and ultimately graduating.

The next step in the process was to reduce the number of variables necessary for predicting first year progression rate. After all of the data sources were identified, the data set had thousands of variables available. Through a process of eliminating variables with a high number of missing observations, variables highly correlated with each other, and making sure a meaningful level of variables were included, the data set was left with 927 variables for 58 North Carolina Community Colleges.

Identifying the right machine learning algorithm to perform an analysis on first year progression was the next step in the process. Many attempts were made with various models, however, an elastic net linear regression model was determined to be the best model for this application of the data set. The number of variables was further reduced to 18 variables using statistical methods for determining importance to the elastic net linear regression model.

Finally, the model was able to identify the most important features for determining first year progression. The top feature was the weighted average of the percentage of

black female high school students in the county of the community college. This feature has a negative correlation to first year progression, meaning that the higher the number of black female high school students in the county of the Community College, the lower the first year progression rate. The next most important feature was the percentage of high beginning English as a Second Language (ESL) students receiving Adult Education and Family Literacy Act (AEFLA) services who achieve an Educational Functioning Level gain during the program year. This feature has a positive correlation to first year progression. This indicates that if there are a high number of students in this AEFLA program that achieve an Educational Functioning Level gain during the program year at the community college, there is a positive impact on first year progression rate. The third most important feature also relates to an Adult Education and Family Literacy Act service. This feature was the percentage of Low Intermediate Basic Education students receive AEFLA services who passed an exam administered by a recognized agency to achieve high school equivalency. The higher the number of students in this program that achieve high school equivalency status, the more positive an impact on first year progression rate. The AEFLA service features included as most important features for First Year Progression makes this research promising. First, this shows that these services are indeed effective in driving first year progression success. Second, the results provide specific areas where more attention and funding can be place to drive success in first year progression even further.

In Section 2, a background of current research on postsecondary educational attainment is presented, with a closer look at how socioeconomic status and background relates. Section 3 discusses the data collection process and more detail about the sources of data identified for the research presented here. In Section 4, initial insights into the data collected are introduced and provide a preview into the findings of the research. An explanation of the elastic net linear regression model that was used to identify the most important features related to first year progression is provided in Section 5. Section 6 provides more detail about the most important features related to first year progression identified by the elastic net linear regression model. In Section 7, General Data Protection Regulation (GDPR) and the overall and ethical implications on research related to higher education is discussed. Section 8 details the conclusions of the research.

2 Background on Postsecondary Educational Attainment

With the growing number of jobs requiring postsecondary education, it is critical that education remains accessible and achievable for students of all backgrounds and socioeconomic statuses. By 2020, it is predicted that an average of 65% of all jobs nationwide will require some form of postsecondary education, including certifications, associate degrees, bachelor's degrees, and graduate degrees [3]. Even more concerning is that the number of states with an average higher than the nationwide average is predicted to increase from 19 to 27 states from 2018 to 2020 [3]. North Carolina is among these states with an estimated 67% of jobs requiring postsecondary education by 2020 [3]. This increase in total percentage of jobs that

require postsecondary education places more weight on schools to ensure their students are prepared and qualified for the workplace.

Currently, the number of students who are completing postsecondary education, whether certificates or degrees, does not surpass or even match the need. According to a report conducted by Georgetown University, the percent of employees who meet or exceed the education requirements for job openings in the state of North Carolina are only around 58% which is slightly lower than the national average of 60% [3]. In North Carolina, there are 406,000 job openings requiring some college, 176,000 requiring an associate degree, 312,000 requiring a bachelor's degree, and 157,000 requiring a master's degree [3]. When looking at community college enrollment, while 83% of North Carolina high school graduates enroll in some sort of postsecondary education, only 52% of those students end up graduating [5] and 49% are considered successful after three years, which refers to either graduating, transferring to a four year college, or are still enrolled [6]. For the rest of the students, while 23% of students leave school during their first year, this percentage declines in subsequent years with only 14% and 8% of students leaving during the second and third years respectively [6].

Students are also taking longer than anticipated to earn degrees at two-year colleges. Only 30% of full-time students at community colleges graduate with a certification or associate degree within three years, while 60% of full-time students at four-year colleges earn a degree within six years of enrolling [7]. It can be speculated that the causes for these differences include working status, socioeconomic status, and family dynamics between students who attend two-year versus those that attend four-year universities. Students who either attend community colleges or are economically disadvantaged have a particularly low college completion rate [7]. These students typically have jobs and family commitments, are non-traditional, or only attend school part-time [7]. Several interventions have been put into place at trial schools including developmental education, student success courses, stronger advising, and connection to communities, but few of these have proved to be successful [7].

Many students who attend community colleges tend to be lower income and first-generation students, may not be prepared for college academically, and often lack connection to the college [8]. Strong faculty connections and academic advising may be a factor that can contribute to the percentage of low-income and first-generation students who stay in school and graduate on time [8]. Most of these students do not receive the academic and social support that is critical for success. In fact, higher income students were shown to be six times more likely to attend and succeed in postsecondary education than low-income students [5]. In the 35 years since this study, this gap has nearly doubled [5]. Additionally, first generation college students are generally less prepared for the transition from high school to college than students with family members who have previously attended college [9].

Certain minority students are also not prepared for this transition even though on average more minority students in the United States attend community colleges than majority students [10]. Nationally, Hispanic enrollment at two-year colleges is at 58%, with just 42% of white students enrolled at these types of colleges [10]. Most students have the intent to transfer to a four-year institution, but less than 25% end up following through on this intention [10].

While there is no single reason that students drop out of community colleges, we aim to identify factors that hold more influence than others. Identifying these factors that could increase first year progression and subsequently graduation rates from North Carolina community colleges can help ensure that workers are qualified and academically prepared for the jobs of the future.

3 Data Collection

First year student success and progression was measured using data from five publicly available locations. These locations included the National Center for Education Statistics (NCES)[1], the North Carolina Community College's page for Analytics and Reporting[2], the North Carolina Office of State Budget and Management (OSBM)[3], the Internal Revenue Service (IRS)[4], and The Belk Endowment Educational Attainment Data Repository for North Carolina Public Schools[5]. Together, these data sources provided a broad range of variables and perspectives for retention, demographics, and other useful information in North Carolina Community Colleges. All of the data collected for this analysis was from the 2016-2017 school year to ensure for completeness and consistency.

The NCES's Integrated Postsecondary Education Data System (IPEDS) contained data on the 58 community colleges related to student enrollment, curriculum completion, retention rates, graduation rates, and student population. Data obtained from the North Carolina Community College page consisted of information on Adult Education and Family Literacy Act (AEFLA) services which provide educational services to students identified as basic skills deficient or English as a Second Language (ESL) students [17]. Performance measures of basic skills student progress,

[1] Use the Data. (n.d.). Retrieved October 12, 2018, from https://nces.ed.gov/ipeds/use-the data

[2] Analytics and Reporting. (2016, April 13). Retrieved October 12, 2018, from http://www.nccommunitycolleges.edu/analytics

[3] North Carolina Office of State Budget and Management. (n.d.). Retrieved October 12, 2018, from http://www.osbm.nc.gov/facts-figures/nc-census-data/nc-census-lookup/decennial-census-occupation-reports-north-carolina-2000

[4] SOI Tax Stats - Individual Income Tax Statistics - 2016 ZIP Code Data (SOI). (n.d.). Retrieved October 12, 2018, from http://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi

[5] Drew J., The Belk Endowment Educational Attainment Data Repository for North Carolina Public Schools, (2018), GitHub repository, https://github.com/jakemdrew/EducationDataNC

student success rate in college-level English and math courses, first year progression, curriculum student completion, licensure and certification passing rate, and college transfer performance are also detailed on this site.

Additionally, information about the communities in the areas surrounding the colleges was added to the data collection from the OSBM, the IRS, and The Belk Endowment Educational Attainment Data Repository for North Carolina Public Schools. The OSBM detailed occupation data for surrounding community members while the IRS provided income data for families in the areas that are likely to attend these community colleges. The Belk Endowment Educational Attainment Data Repository for North Carolina Public Schools included local public high school data related to areas such as school performance, student demographics, district funding, and teacher experience. This data was consolidated using a weighted average based on the number of students in the school and matched to the community college closest in geographical distance.

The aforementioned data sets were merged together based on college and county name to generate one data set describing the community colleges with 1950 attributes. After removing unnecessary, highly correlated, and missing data, the final data set specified 927 variables related to North Carolina community colleges and the surrounding communities. The various data sources used for this analysis created a holistic data set by contributing unique perspectives regarding community colleges in North Carolina. This can be used to investigate factors related to student retention in the form of first year success.

4 North Carolina Community College Retention

As of 2018, there were 102 counties in North Carolina with only 58 community colleges representing those counties. A total of 44 counties in North Carolina have no community college. This represents 293,000 students or 23% of the population of students in the public school system (kindergarten through high school) in 2017 without a community college nearby. Twenty three percent of students in the North Carolina public school system have a proximity barrier to consider when attending community college in North Carolina. Figure 1 shows a map of the counties in North Carolina with a community college by student size. The size of the student population in a county is represented by color, with darker colors implying a higher number of students. As shown in Figure 1, there are 44 counties with no color, which means that county does not have a community college.

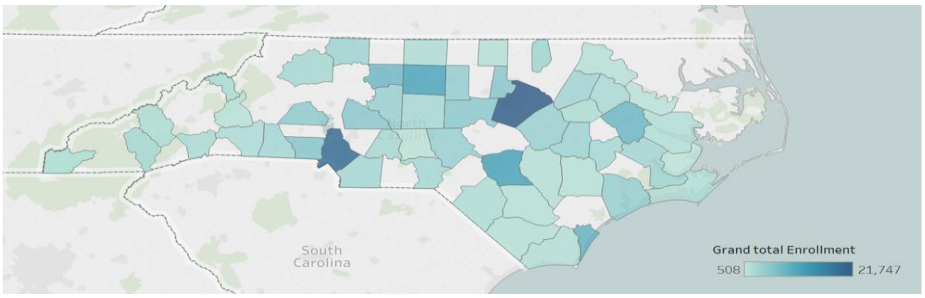

Figure 1. Map of North Carolina Community Colleges by County and Size

Overall, looking at 2016 student population demographics, there is a 5% decrease in minority representation in the community colleges related to the representation in kindergarten through high school. In 21 out of the 55 community college, there is a more than 10% decrease in minority representation. An example is in Vance county, where Vance-Granville Community College resides. In the secondary school system there is an 81% minority representation. In contrast, at the community college there is only a 49% representation, a decrease of 32%. This decrease could signal an opportunity to try to increase the number of minority students in these areas that attend community colleges. This decrease is shown in Figure 2.

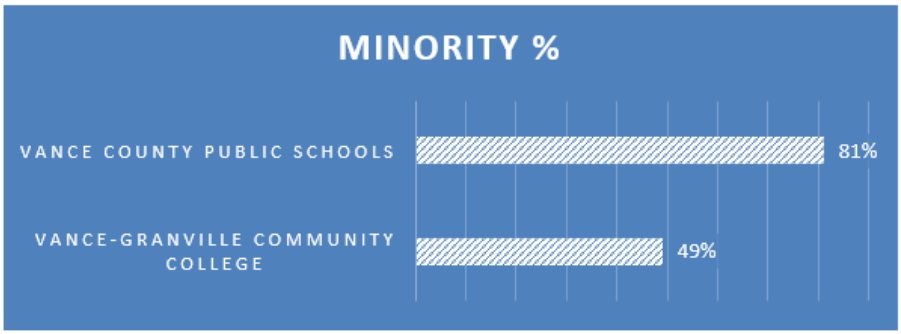

Figure 2. Minority Percentage Representation Public Schools versus Community College - Vance County

While research shows economic factors are a possible reason for North Carolina students not returning in their second year, only 2% - 16% of these students receive financial aid of any kind [2]. Only 3 community colleges in North Carolina have 10% of students receiving financial aid. This presents an opportunity to provide more assistance to financially deserving students attending these institutions. Figure 3 shows the top ten schools with the highest percentage of students receiving financial aid of any kind.

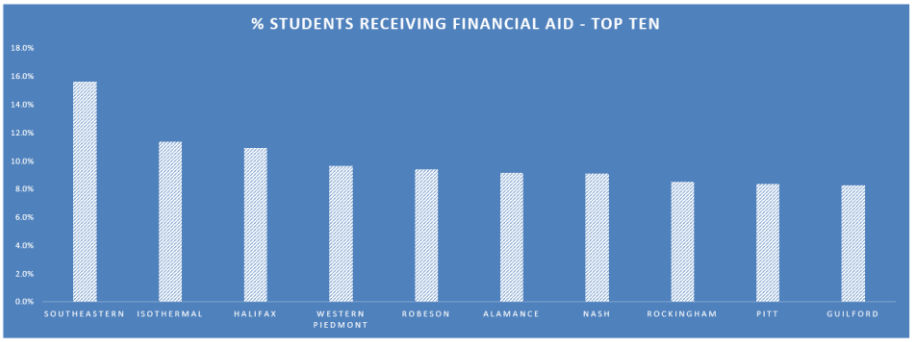

Figure 3. Percentage of Students Receiving Financial Aid by Community College

Another interesting metric to note is that the number of women students ranges from 53% to 76%, with 35 of the 54 community colleges having more than 60% women. The overall average of women community college students is 60%, with men representing 40%. However women represent only 55% of the community college students that successfully meet the criteria for first year progression. This can be seen in the charts in Figure 4.

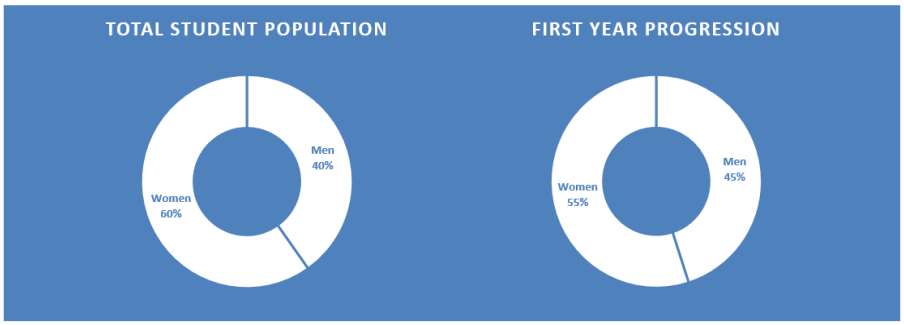

Figure 4. Student Population versus First Year Progression Population – Sex

While interesting statistics and possible areas of focus have been identified, attention is turned toward First Year Progression. First year progression for the North Carolina community colleges is on average around 70%, with a range of 54% to 81%. Caucasian males and females, Hispanic females, and other minority females are all above the 70% average for first year progression rate. Hispanic males and other minority males are at about the 70% average. African-American females (61%) and African-American males (51%) fall well below the 70% average. From an age perspective, all age groups except for students 19-24 are at or above the 70% average, however the 19-24 age group, representing those most recently out of high school, have only a 60% first year progression.

Figure 5 shows that while the overall population of black or African American students is about 23%. However when looking at the population of successful first year progression students, black or African Americans represent only about 16% of the students. The other race groups either increase or stay the same overall.

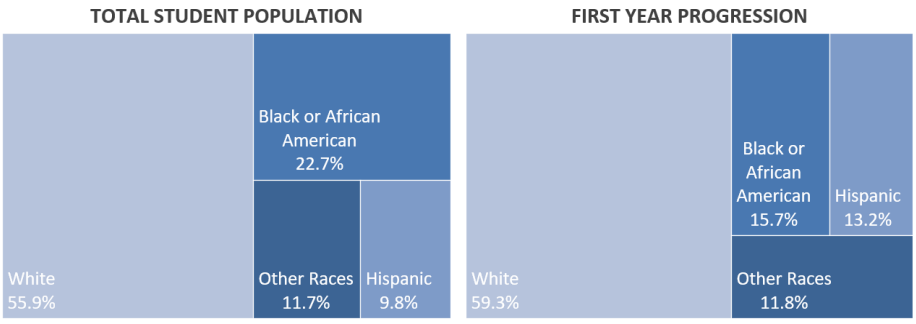

Figure 5. Student Population versus First Year Progression Population - Race

Initial insights into the North Carolina community college data provide areas to keep in mind in the deeper analysis that will follow. Distance, sex, and race factors could become important in trying to identify the most important features in determining first year progression rates.

5 Model Creation

5.1 Elastic Net Linear Regression

A linear regression model was determined to be to best choice due to the fact that the target variable, first year progression, was a continuous variable. Multiple modeling and variable selection techniques were attempted but the best performing technique was elastic net variable selection. Elastic net works by assigning each variable a penalty term [23]. These penalty terms are then combined to find a loss function for each combination, with the ideal combination being the one with the smallest loss function [23]. The penalty term used in elastic net, figure 6, is a combination of the penalty terms from lasso and ridge variable selection [23].

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|\right),$$

Figure 6. Elastic Net penalty term

In this formula alpha has a range between 0, ridge, and 1, lasso, while also tuning lambda [23]. Using elastic net there are two variables to tune as opposed to lasso and ridge selection where there is only one variable. This allows it to be a hybrid of the other two variable selection techniques, and thus mitigate their shortcomings. Lasso selection does not perform well when there are highly correlated variables in the dataset and ridge selection does not perform well when there are many noisy

variables. Elastic net performs well in both of these situations, which made it the ideal choice in this case.

5.2 Model Execution and Results

A 10-fold cross validation with a train test split of 90/10 was used when creating the model. This split was chosen due to the small sample size in the dataset to ensure there was enough training data to support the model building phase. In the full data set there were 927 variables, which decreased to 894 variables after removing those that were variations of first year progression rate. This subset was used as the modeling data set, however the elastic net variable selection only selected 18 variables to use in the linear regression model. These variables, listed in table 1, had a coefficient that did not equal zero whereas the other variables did.

Table 1. Variables with absolute weights above zero from model

Variable	Description
BlackFemalePct	Weighted average of the percentage of black female high school students in the county.
HighBeg_ESL_PCTProgress	Percentage of High Beginning English as a Second Language students receiving AEFLA services who achieve an Educational Functioning Level gain during the program year.
LowIntBasicEd_HSE	Percentage of Low Intermediate Basic Education students receiving AEFLA services who passed an exam administered by a recognized agency to achieve high school equivalency.
LowIntBasicEd_AHSGrad	Percentage of Low Intermediate Basic Education students receiving AEFLA services who graduated with a secondary school diploma.
MEEL	Number of times a college met or exceeded the excellence level in the categories of Basic Skills Progress, Credit English Success, Credit Math Success, First Year Progression, Curriculum Completion Rate, Licensure Passing Rate, and Transfer Performance.
MinorityMalePct	Weighted average of the percentage of minority male high school students in the county.
Blet_PCTPassing2016	Percent of students who passed the NC Department of Justice, Criminal Justice Standards Division's Basic Law Enforcement Training (BLET) Exam in 2016.
HighIntBasicEd_HSE	Percentage of High Intermediate Basic Education students receiving AEFLA services who passed an exam administered by a recognized agency to achieve high school equivalency.
EMTParamedic_PCTPassing2017	Number of students who passed NC Office of Emergency Medical Services' EMT Exam.
MHL_Hispanic_ENROLL_sch_pct	Weighted average of the enrollment percent of Hispanic male high school students in the county.
EOCEnglish2_GLP_SWD	Weighted average of the high school students with disabilities who achieved grade level proficiency on the EOC English 2 test in the county.
EOG/EOCSubjects_GLP_LEP	Weighted average of the high school students with limited english proficiency who achieved grade level proficiency on the EOG/EOC subjects tests in the county.
American Indian or Alaska Native total	Number of American Indian or Alaska Native students at the community college.

GraduationRate_5yr_Female	Weighted average of the 5-year high school graduation rate for females in the county.
lea_short_susp_per_c_num	Weighted average of the number of short term suspensions per 100 students at school level in the district.
AwardLess2Year_Women_Number	Total number of women who have completed less than two years but more than one year of coursework.
ACTSubtests_BenchmarksMet_Asian	Weighted average of the benchmarks met on the ACT subtests by Asian high school students in the district.
EOCSubjects_GLP_TwoorMoreRaces	Weighted average of high school students of two or more races who achieved a grade level proficiency on the EOC subjects test in the district.

The model had a R^2 of 73.5%, meaning that nearly three quarters of the variance in first year progression rate is explained by the model. Furthermore, the average fold had a mean absolute error (MAE) of 0.0408, a root mean squared error (RMSE) of 0.0502, and a mean absolute percentage error (MAPE) of 6.10%. These statistics are encouraging considering the range of first year progression rate is 54% to 85% which is a range of 31 percentage points. Taking one of these statistics as an example, a MAE of 0.0408 means that on average the model's predictions was only off about 4 percentage points. As seen in figure 7, the predicted values compared to the actual values are reasonably close to the ideal line. Observations at the beginning and end of the range are farther away from the line than observations in the middle, which is to be expected. The selection of 18 variables in the model out of nearly 900 provides evidence that if schools focus on improving just a few characteristics within the college they can increase their first year progression rate.

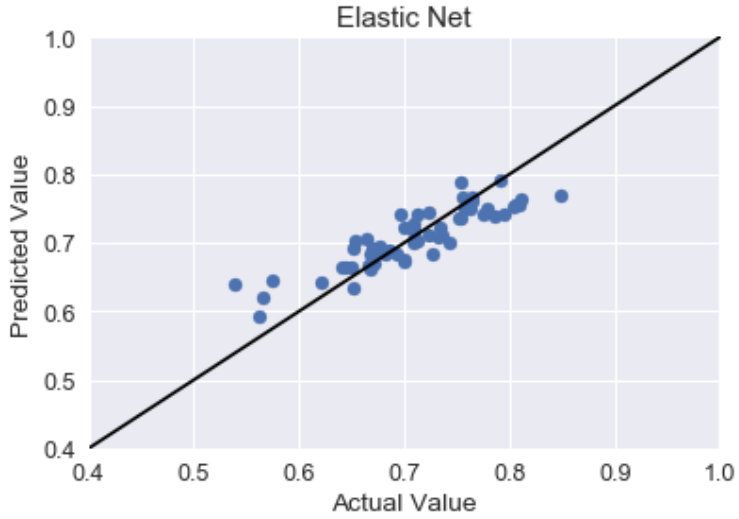

Figure 7. Predicted vs. Actual plot from model

6 Identifying Leading Factors for Postsecondary Attainment

A feature importance analysis was conducted to determine the factors that are the biggest predictors for first year progression. Table 2 displays the top five features that have the largest influence when predicting first year progression rate. The features are accompanied by a brief description of the variable along with the weight the variable has on the model. The feature importance weights are the variable coefficients in the regression equation.

Table 2. Mean feature importance of the top five variables used to predict first year progression rate

Variable	Description	Feature Importance
BlackFemalePct	Weighted average of the percentage of black female high school students in the county.	-0.1095
HighBeg_ESL_PCTProgress	Percentage of High Beginning English as a Second Language students receiving AEFLA services who achieve an Educational Functioning Level gain during the program year.	0.0632
LowIntBasicEd_HSE	Percentage of Low Intermediate Basic Education students receiving AEFLA services who passed an exam administered by a recognized agency to achieve high school equivalency.	0.0589
LowIntBasicEd_AHSGrad	Percentage of Low Intermediate Basic Education students receiving AEFLA services who graduated with a secondary school diploma.	0.0406
MEEL	Number of times a college met or exceeded the excellence level in the categories of Basic Skills Progress, Credit English Success, Credit Math Success, First Year Progression, Curriculum Completion Rate, Licensure Passing Rate, and Transfer Performance.	0.0211

The variable with the largest influence on first year progression rate was the percentage of black female high school students in the county. A coefficient weight of -0.1095 suggests that every 1% increase in the percentage of black female students in the county is associated with an estimated 10.95 percentage point decrease in predicted first year progression rate holding all other factors constant. Figure 8 displays the correlation between black female percent and first year progression rate, revealing an R^2 value of -0.51 (p-value = 0.000041). This indicates that 51% of the variance in the first year progression rate can be explained by the percentage of black female students in the county.

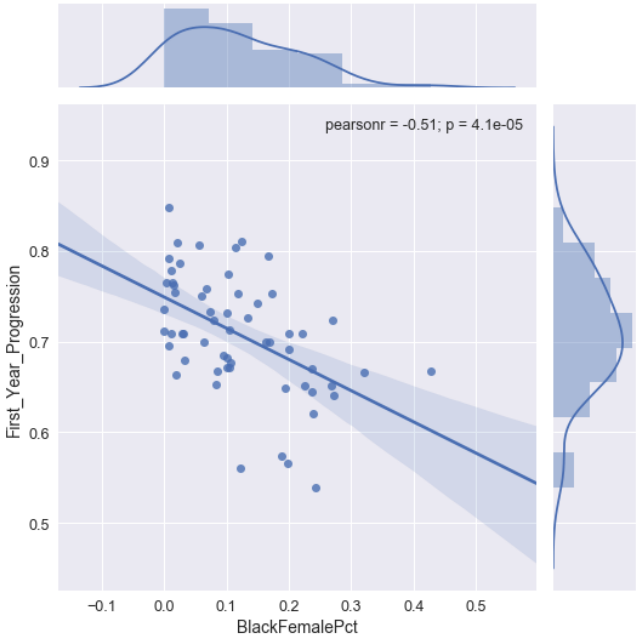

Figure 8. Correlation between the percentage of black female students in the county and first year progression rate

The second, third, and fourth most important variables related to first year progression rate are all associated with specific levels of the AEFLA program, which provides individuals Adult Education and Family Literacy Act (AEFLA) services during the program year. Students that receive this program are basic skills deficient, have not achieved a high school equivalent level of education, or are English language learners [17]. Each of these variables estimates an increase in first year progression rate, implying that these services seem to have a positive impact on educational attainment in North Carolina. Every percentage increase of High Beginning English as a Second Language students receiving these services who achieve an Educational Functioning Level gain during the program year is associated with a 6.32 percentage point increase in the estimated first year progression rate value. An Educational Functioning Level gain is defined as an increase in program level according to state guidelines based on a pre and post test score in reading, writing, listening, or math [21]. The Educational Functioning Levels are as follows [21]:

- Adult Basic Education (ABE)
    - Beginning ABE Literacy
    - Beginning Basic Education
    - Low Intermediate Basic Education
    - High Intermediate Basic Education
    - Low Adult Secondary Education
    - High Adult Secondary Education
- English as a Second Language (ESL)

- Beginning ESL Literacy
- Low Beginning ESL
- High Beginning ESL
- Low Intermediate ESL
- High Intermediate ESL
- Advanced ESL

These levels are monitored by the College and Career Readiness Unit of the North Carolina Community Colleges' Programs and Student Services Division. This unit allocates federal and state funds to eligible providers and provides leadership, professional development opportunities, and technical aid to these providers [22]. In this case, the Educational Functioning Level gains are reported by each community college based on the services they provide that comply with the federal regulations to receive these grants. The services provided by the community college to high beginning ESL students and low intermediate basic education students are associated with higher first year progression rates.

The importance of each variable can also be interpreted by looking at the effect that each feature has on the predictiveness and quality of the model. The original model had an R^2 value of 73.5%. If the most important variable, the average percentage of black female high school students, is removed from the model, the value of R^2 decreases by 11.4 percentage points to 62.1%. Additionally, removing the second most important variable decreases the value of R^2 by 60.4 percentage points to 13.1%. These variables as well as the relationship between the variables used to build the model affect the variability of the model.

7 Ethical Implications

In April 2016, the European Parliament and Council agreed upon the General Data Protection Regulation (GDPR) [18]. At its core, GDPR "mandates a baseline set of standards for companies that handle EU citizens' data to better safeguard the processing of movement of citizens' personal data" [18]. This includes:

- "Requiring the consent of subjects for data processing
- Anonymizing collected data to protect privacy
- Providing data breach notifications
- Safely handling the transfer of data across borders
- Requiring certain companies to appoint a data protection officer to oversee GDPR compliance" [18]

While this might sound like it only pertains to European Union (E.U.) members, compliance with regulations is required by any company that "markets goods or services to EU residents, regardless of its location" [18]. This would include United States institutions of higher education. "Like their counterparts in the business world, U.S. colleges and universities are scrambling to figure out how the rules apply to their overseas programs as well as the data they collect on students and employees who are E.U. citizens" [19].

There are many ethical and legal debates that take place around GDPR, including "the right to be forgotten", which, if requested a citizen can have any data on the

internet pertaining to them erased [20]. This may include links to negative articles that are old enough that they are deemed irrelevant, which could be seen as a violation of freedom of press in the U.S. [20]. There are also arguments being made for the U.S. to adopt similar GDPR laws to assist in protecting the privacy of U.S. citizens [20]. All of these regulations and potential trends should be watched by data scientists to anticipate and even react to required changes in the future. For the scope of this paper, the topic will be narrowed to U.S. higher education data as it relates to E.U. citizens and the ethical implications for data scientists.

Data has become increasingly readily available from various sources, especially the Internet. For companies, as well as universities and colleges to become compliant with GDPR regulations, it is not a trivial endeavor. Data containing information related to E.U. citizens has likely been made available prior to GDPR going into effect. While legally, it is the responsibility of the entities to become compliant with GDPR, data scientists have an ethical obligation to take care with any data used for research purposes. This obligation is not limited to E.U. citizen's data when it comes to using personal information that can be linked back to an individual person. Ethically, a data scientist should notify the source that personal information is being made available or, in the case of GDPR, notify the source that they are not in compliance with the regulations.

As mentioned, the Internet has become a primary source for identifying data sets that can be used to further research. As data scientists use the web to obtain datasets through the means of web scraping, there are steps that should be taken to ensure that an ethical approach is maintained. "Web scraping is a term for various methods used to collect information from across the Internet" [14]. It has become increasingly easy with a few lines of code in Python or R, that all the information from a website can be taken from the page and used for analysis. Researchers must consider what the data is being used for and make sure that it is being used for responsible purposes [15]. There are several approaches that researchers should follow when web scraping. One approach is to include a "User Agent" string in the code that identifies the researcher, their intention for the data, and a method of contact that gives the website owner information about the scraping that is happening on their site [15]. This would also assist entities in maintaining compliance with GDPR by being able to reach back out to the researchers or just for awareness for how widespread a data leak has become. Another is for the researcher to always give credit for where the data was obtained when using it as part of a post, article, or submission of any kind [15]. Researchers should also use public API's for the information, if available, and not use a scraping method instead [15]. Only relevant data to the purpose of the researcher should be kept from the website, the rest should be discarded [15]. Data scientists should take care to anonymize any personal information that they may come across and notify the data owner of the possible leak of personal information that has made its way to Internet. Finally, the purpose of the scraping should be to create "new value from the data, not to duplicate it" [15]. If data scientists and researchers maintain these approaches, they can ensure there is no issue from data owners with the data being used for the purposes intended and maintain an ethical approach to scraping the web for additional information that may assist in furthering the research conducted. All of this will help data owners to maintain compliance with GDPR.

The shift in mindset in the U.S. toward more privacy around personal data, and considering the impact of GDPR from a regulatory perspective, there could be implications on the type of research done in this paper. While it seems that there is a need for more data in order to make proper decisions on policy and changes for the benefit of community college students, it cannot be at the expense of any student's privacy. Future research in this area must be done with care and an ethical framework in mind. Anonymization and masking is done to data where there are few students making up the results. This is done to protect the privacy of those students. It may be easy for data scientists to derive the true numbers based on other numbers, but this should not be done. Data scientists must respect the attempt at keeping this data private, use a general method of imputation, or risk having the data not available at all. Data scientists must also keep in mind that in future iterations of the research done on community colleges in North Carolina, that there may be less data available due to the drive toward more privacy and college's obligation to adhere to GDPR regulations. This should not cause the abandonment of this important research, but challenge data scientists to be more diligent in their efforts to derive insights from these datasets while adhering to an ethical framework for working with the data.

8 Conclusions

The community college system in North Carolina is a developing system that can be utilized to help students get the education they both need and desire. Success in a community college can be quantified in multiple ways; some examples are transferring to a four-year degree, acquiring an associate's degree, and staying on-track in terms of credits. Looking at first year retention as our target variable, we were able to determine some key features. An elastic net linear regression model revealed that the three most influential features are the percentage of black female high school students in a particular county, the percentage of English as a Second Language (ESL) students who receive AEFLA services who increase in their educational function level, and the percentage of low intermediate basic education students who receive AEFLA services who passed a high school equivalency exam. This alone implies two important things. First, that as the percentage of black high school females increases in a county, first year progression rates of the community college in that county tends to decrease. Secondly, the results indicate that as students succeed in an AEFLA services, the first year progression rates of a particular community college tend to increase.

While the information gathered is intriguing, it does not help the system unless there is the motivation and ability to make changes in the state of North Carolina. Stagnant knowledge is useless without a plan moving forward. Our hope with this particular knowledge is to make it available to future students and researchers so that they may apply their own models to the data, improve the models we already have in place, or use the models as they are to start putting some changes into place. It is important to note that since this analysis was completed on all of the community colleges in North Carolina, the results of this study cannot be generalized to other states or colleges. This analysis along with future analyses can help North Carolina

community colleges improve retention and help all students achieve academic success.

References

1. Endowment, B. (n.d.). What We Fund. Retrieved September 20, 2018, from http://jmbendowment.org/what-we-fund/
2. Carolina Demography. (2018). North Carolina's Leaky Educational Pipeline & Pathways to 60% Postsecondary Attainment [Word Document].
3. Carnevale, Anthony P., Nicole Smith, and Jeff Strohl. (2013). Recovery: Job Growth and Education Requirements Through 2020 – State Report. Washington, DC: Georgetown University Public Policy Institute
4. Poole, S., King, J., Bullis, J., Burch, F., & Peal, J. (n.d.). Analysis of The North Carolina High School to Community College Articulation Agreement's Impact on Student Motivation in a North Carolina High School. ProQuest Dissertations Publishing. Retrieved from http://search.proquest.com/docview/873816057/
5. Sanchez, L., Buss, R., Gonzales, S., & Span, D. (n.d.). Preparing High School Students for Transition to Community College. ProQuest Dissertations Publishing. Retrieved from http://search.proquest.com/docview/1901472385/
6. Horn, Laura, and Thomas Weko. 2009. On Track to Complete? A Taxonomy of Beginning Community College Students and Their Outcomes 3 Years After Enrolling: 2003-04 Through 2006, Statistical Analysis Report. Washington, DC: U.S. Department of Education. http://nces.ed.gov/pubs2009/2009152.pdf
7. Karp, M., & Bork, R. (n.d.). "They Never Told Me What to Expect, So I Didn't Know What to Do": Defining and Clarifying the Role of a Community College Student. Teachers College Record, 116(5).
8. Mcarthur, R. (n.d.). Faculty-Based Advising: An Important Factor in Community College Retention. Community College Review, 32(4), 1–I. Retrieved from http://search.proquest.com/docview/213286778/
9. Hockersmith, W., Swayze, S., Stephenson, K., & Tekleselassie, A. (n.d.). Transition Experiences of First-Generation Students Enrolled in a High School to Community College Partnership Program. ProQuest Dissertations Publishing. Retrieved from http://search.proquest.com/docview/1778511110/
10. Crisp, G., & Nora, A. (n.d.). Hispanic Student Success: Factors Influencing the Persistence and Transfer Decisions of Latino Community College Students Enrolled in Developmental Education. Research in Higher Education, 51(2), 175–194. doi:10.1007/s11162-009-9151-x
11. Endowment, B. (n.d.). Operating Principles. Retrieved October 30, 2018, from http://jmbendowment.org/who-we-are/operating-principles/
12. Endowment, B. (n.d.). Frequently Asked Questions. Retrieved October 30, 2018, from http://jmbendowment.org/frequently-asked-questions/
13. Endowment, B. (n.d.). Funding Criteria. Retrieved October 30, 2018, from http://jmbendowment.org/how-we-work/funding-criteria/
14. Web Scraping. (n.d.). Retrieved October 30, 2018, from https://www.techopedia.com/definition/5212/web-scraping
15. Densmore, J. (2017, July 23). Ethics in Web Scraping – Towards Data Science. Retrieved October 30, 2018, from https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01
16. Tracking Students to 200 Percent of Normal Time: Effect on Institutional Graduation Rates. (2010). U.S. DEPARTMENT OF EDUCATION. Retrieved January 25, 2019, from https://nces.ed.gov/pubs2011/2011221.pdf

17. Electronic Code of Federal Regulations. (n.d.). Retrieved March 10, 2019, from https://www.ecfr.gov/cgi-bin/text-idx?SID=f284d9f4d8e105d9b4ccbc1dfc7c5fc6&mc=true&node=pt34.3.463&rgn=div5#se34.3.463_123
18. De Groot, J. (2019, January 3). What is the General Data Protection Regulation? Understanding & Complying with GDPR Requirements in 2019. Retrieved March 9, 2019, from https://digitalguardian.com/blog/what-gdpr-general-data-protection-regulation-understanding-and-complying-gdpr-data-protection
19. Raths, D. (2018, May 24). What GDPR Means for U.S. Higher Education. Retrieved March 9, 2019, from https://campustechnology.com/articles/2018/05/24/what-gdpr-means-for-us-higher-education.aspx
20. Berman, S. (2018, May 23). GDPR in the U.S.: Be Careful What You Wish For. Retrieved March 9, 2019, from http://www.govtech.com/analysis/GDPR-in-the-US-Be-Careful-What-You-Wish-For.html
21. North Carolina Title II Adult Education Provider Comprehensive Report on Measurable Skill Gains (Rep.). (2017, December 4). Retrieved March 10, 2019, from North Carolina Community Colleges website: https://www.nccommunitycolleges.edu/sites/default/files/data-warehouse/2018_measurable_skills_gain_report_20181212_final.pdf
22. Spell, S. (2019, February 22). Workforce Innovation and Opportunity Act - Adult Education and Family Literacy Act (WIOA - AEFLA). Retrieved March 10, 2019, from https://www.nccommunitycolleges.edu/college-and-career-readiness/wioa-aefla
23. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society,67, 301-320. Retrieved March 19, 2019, from http://www.recognition.mccme.ru/pub/papers/L1/elasticnet.pdf