Predicting First Year Retention for Undergraduate Educational Opportunity Fund Students

By

Kelly O'Neill, Bachelor of Science in Mathematics

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Applied Mathematics

Spring, 2024

Committee Members:

Dr. Osei Tweneboah, Advisor

Dr. Amanda Beecher, Reader

Dr. Nicole Videla, Reader

COPYRIGHT

Dedication

To all of the women in STEM who have ever doubted themselves or questioned their abilities, you are strong, capable, and intelligent. Know your worth and keep pushing the boundaries!

Acknowledgements

First, thank you to my amazing parents who have supported me from day one. None of this would have been possible without your love, guidance, and patience. Thank you for always believing in me, even when I did not believe in myself, and for always being there for me. I know these past six years have not been easy, but I truly would not be where I am today without you, nor would I be the person I am today without you. I feel like I won the lottery being your daughter and I love you both so much.

To my loving sisters who have been cheering me on from the west coast, I'm so grateful for the relationship we have with each other. Thank you for showing me the beauty of balance, and for your endless life advice. You both have inspired me to explore, be more adventurous, have fun, and to try new things. Thank you for everything, and I love you both so much!

To my strong, incredible, intelligent friends, thank you for always encouraging me and being my cheerleader when the semester seemed impossible. Your friendship has been the best gift, one that I am grateful for every day! Here are all of the good times to come!

To my Ellabelle, you are the cutest, sweetest, most loving, and best doodle I could ever ask for. Thank you for being a great study buddy and for all of the hiking and walking breaks. You have made me the proudest dog-mom in the entire world!

Dr. Beecher, I could not have made it through this program without you. From day one, you have always been there to answer my never-ending list of questions, and you helped me navigate my way through the program. Not only are you incredibly intelligent, about both

mathematics and teacher education, but anyone who has met you can see that you just have an infectious and beautiful personality. You have a warm energy about you, and you always have a smile or joke, that immediately makes anyone talking to you feel comfortable. Regarding this thesis, I could not have done this without you. You kindly connected Dr. Videla and I and listened to all of my ideas at various stages. Finally, without you, I would not be getting certified to teach computer science this summer. Dr. Beecher, I am forever grateful to you!

To Dr. Tweneboah, my amazing thesis advisor. I cannot thank you enough for your help during this process and in the preparation of it. Without your machine learning course, I never could have completed this thesis. I recognize that I required a lot of thesis meetings and asked way too many questions, but thank you for your patience, guidance, support, and encouragement. I feel so lucky to have been your student this semester! You are truly one of the kindest, smartest, and best professors I have ever had at Ramapo!

To Dr. Videla, thank you so much for going out of your way and being so accommodating and enthusiastic about partnering with me for my thesis. I am grateful to you for having provided me with not only your precious time, but your valuable data. Thank you for being so kind, gracious, and helpful throughout this entire process. I hope these results are valuable to you!

Finally, thank you to Ramapo College. Although the past 6 years have not always been easy, I am thankful for the amazing friends I have made, the wonderful professors I have had, and the education I have gained. Hopefully one day you will open up an applied math doctoral program!

Table of Contents

List of Tables

List of Figures

Abstract

Predicting undergraduate retention using various machine learning algorithms has the potential to reduce the likelihood of attrition for students who are identified as being at an elevated risk of dropping out. Thus, providing a mechanism to help increase the likelihood of a student graduating from college. Following the approach of previous studies, retention is predicted using primarily freshman data, where retention is defined as a student being enrolled a year later from their first semester. For this thesis, the population was focused on predicting retention for Educational Opportunity Fund (EOF) students. Based on the EOF department's most recent report, which comes from Ramapo's Office of Institutional Research 2023, in 2016, the 4-year graduation rate is 46.40%, and the 6-year graduation rate is 63.10%, whereas for the college, the four-year graduation rate is 56.9%, and the six-year graduation rate is 69.5%, using the Fall 2018 cohort. Through identifying these specific individuals who will not be retained, it allows the EOF department to devise an appropriate plan and provide resources to help the students achieve academic success, and thus increase graduation rates.

This thesis will consider many factors, provided by the EOF department, from Fall 2013 to Spring 2023. I will consider the impact of covid within my analysis. I predict retention using logistic regression, decision tree, random forest, support vector machine, ensemble, and gradient boosting classifier, where feature selection and the Synthetic Minority Over-sampling Technique (SMOTE), since the dataset was not balanced, were used for each algorithm. While all of the models performed well, even after 10-fold cross-validation, the random forest model using feature selection a balanced dataset is recommended. In the future, the EOF department can use

this model to determine which incoming students are at elevated risk of dropping out and provide them with the necessary resources to help them succeed.

The second part of this thesis is a comprehensive exploratory data analysis to learn more about the EOF student population. EOF students tend to struggle within the subject areas of math, biology, interdisciplinary studies, psychology, and chemistry. More specifically, in the courses math 108, interdisciplinary study 101, biology 221, critical reading and writing 102, amer/intl interdisciplinary 201, math 101, and math 110. Regarding retention, the average cumulative GPA for students who retained was 2.84, and 2.15 for students who did not retain. Furthermore, the average term GPA for those who retained was 2.67 but was 1.65 for students who did not retain.

Through analyzing the relationship between retention and other variables, such as GPA, subject areas, and courses, it provides the EOF department with a better idea of possible support mechanisms for students. Coupling this information with the recommended prediction algorithm of ensemble learning, can help the EOF department increase their four year and six-year graduation rates, by providing the student(s) with resources, guidance, and plans with their expertise.

Introduction

The ability to predict which undergraduate students will not transition from freshman year to sophomore year and provide them with resources to help them continue their college education, and eventually graduate, is beneficial to the individual, the program, the college, and society. The first year of an undergraduate's career is vulnerable since "around half of the dropouts among students of the same cohort occur in the transition from the Freshman college year to Sophomore year" (Dursun et al, p.2, 2023). These statistics are further substantiated by a report from the National Center for Education Statistics that cites that around three quarters of undergraduates nationwide drop out of college after their freshman year (Dursun et al, p.2, 2023). In a 2019 study that involved a dataset comprised of 66,060 students from a public United States university, researchers concluded that "students' second year re-enrollment and eventual graduation can be accurately predicted based on a single year of data" (Aulck et al., pg. 9, 2019).

The impact of student attrition is not an isolated concern as student retention is critical to the survival and longevity of post-secondary education as low-retention rates can have a negative perception and detrimental effect on securing future enrollment. On average, post-secondary institutions lose $16.5 billion dollars annually due to the loss of student tuition payments from attrition (Neal A. Raisman, p.4, 2013). This monetary loss is not isolated as it occurs within both federal and state governments as "the US Department of Education Integrated Postsecondary Education Data System (IPEDS) shows that between 2003 and 2008, state and federal governments together provided more than $9 billion in grants and subsidies to students who did not return to the institution where they were enrolled for a second year" (Sandra C. Matz et al., p.1, 2023). Overall, an increase in retention rates could help stabilize the future of the institution

as it would be more appealing to “the legislators and policymakers who oversee higher education and allocate funds, the parents who pay for their children’s education in order to prepare them for a better future, and the students who make college choices [looking] for evidence of institutional quality and reputation to guide their decision-making processes” (Dursun Delen, p. 498, 2010). Furthermore, it is primarily through these attrition rates, amongst other statistics reflecting the extent of their educational efficacy, that colleges receive funding opportunities and government aid (Lovenoor Aulck et. al., p.9, 2019).

Post-secondary education is a time intensive commitment, and for many students, a financially stressful pursuit. It is estimated that over 65% of undergraduate students in the United States receive some type of student loan throughout the duration of their studies (Sandra C. Matz. et al, p.1, 2023). “students who drop out of university without a degree, earn 66% less than university graduates with a bachelor’s degree and are far more likely to be unemployed” (Sandra C. Matz et al., p.1, 2023). This lack of unemployment opportunity as a result of their limited postsecondary educational attainment typically leads students “to head down a path that leads to lower-paying jobs, poorer health, and the possible continuation of a cycle of poverty that creates immense challenges for families, neighborhoods, and communities” (Mohammad Arif Ul Alam, p.1, 2021). The inability to obtain a bachelor’s degree limits employment opportunities for students, thus putting them at a financial disadvantage and compounding monetary stress due to the statistically increased likelihood of having difficulty repaying their loans. In conjunction to their financial difficulties associated with their attrition, students may have to navigate the adverse consequences, such as feelings of disappointment or failure, to their mental health (Sandra C. Matz et al., p.1, 2023). However, the availability and subsequent distribution of grants and funding can help alleviate a portion of financial stress for students, but it should serve

as a data-driven decision to increase the allocation of funds to certain types of financial aid" (Alexandre M. Ohlbrecht et al., p.4, 2016).

There are many factors that must be considered, such as socioeconomic background. Undergraduate students who are under financial stress or identify as being first-generation students are more likely to discontinue their studies, and not graduate, compared to undergraduate students who receive financial aid either through scholarships or family, for example, which does not have to be paid back (Sandra C. Matz et al., p.2, 2023). Within a six-year time period, first generation and students with a lower socioeconomic status are less likely to earn their bachelor's degree as "among high school sophomores whose parents were in the lowest income group in 2001, 21% of those who earned at least a bachelor's degree, 17% of those with an associate degree, and 13% of those with only a high school diploma had reached the highest income quartile themselves 10 years later" (CollegeBoard). So, even if students are classified as being low-income, there is an increased chance for them to break the cycle of poverty and move up the financial ladder by earning their bachelor's degree.

The ability to predict student retention based on first year data is critical, as 'the earlier one can identify students who might struggle, the better the chances that interventions aimed at protecting them from gradually falling behind and eventually discontinuing their studies will be effective" (Sandra Matz et al., p.2, 2023). For this thesis I examine retention in a more focused subset of the student population. I will consider those who are a part of the Educational Opportunity Fund (EOF) at Ramapo College instead of trying to predict retention for all undergraduate students of varying backgrounds and socioeconomic status. The EOF department's purpose is to "[provide] meaningful access to higher education for qualified New Jersey students impacted by historical poverty" (PURPOSE OF EOF). In addition to meeting the

financial eligibility requirements, students must demonstrate high motivation through having an academic average equivalent to a B- or higher. Once students are accepted in the program, they will work with student development specialists who interweave the students' individual goals with their academic ones. These efforts aim to proactively prepare students for academic and social challenges students will face during their collegiate career.

In order to receive any financial aid, students must file the Free Application for Federal Student Aid, (FAFSA). Additionally, within the program there are multiple opportunities for students to receive supplemental financial assistance. "On average, over one million dollars in Ramapo College Grants (RCG/EOF) are annually awarded to students enrolled" in the EOF program (EOF). In the Ramapo College EOF program, students entering their freshman class have the unique opportunity to acclimate to the college lifestyle in the summer and take courses early. Through a combination of EOF state and college grants, students are able to take a maximum of six credits, and stay in their assigned first-year housing, at no cost to them.

There are many approaches for predicting retention within the literature, but they share two features; using the abundance of readily available data collected by the college, and the implementation of machine learning algorithms. Within this paper I use several machine learning algorithms – support vector machine, logistic regression, decision tree, random forest, ensemble learning, artificial neural networks, and a gradient boosting classifier- to predict retention, of the first-year Ramapo College EOF student using college data from 2013 through 2023. The primary aim of this study is to identify students who are likely to drop out of college, so that the EOF department can appropriately allocate resources and develop a plan to help the student persist and eventually graduate. (Ruba Alkhasawneh et al, pg. 35, 2014). Since, "the earlier one can identify students who might struggle, the better the chances that interventions aimed at protecting

them from gradually falling behind- and discontinuing their studies- will be effective" (Sandra C. Matz et al., pg.2, 2023). The secondary aim of this study is to look at any trends in the demographics of students who have and have not retained. I offer an additional focus on STEM, where it is known to have significant major attrition issues. I will examine which classes have historically challenged EOF students academically.

Chapter 1: Background

Studying retention, and therefore its complement attrition, has long been both an interest and concern to the respective institution, which is reflected by the extensive body of literature detailing various methodologies, time spans, perspectives, and conclusions as detailed in the *Introduction*. Since "colleges and universities collect vigorous amounts of data from students as soon as they apply to the institution," all of the subsequent studies described use data-based approaches by implementing machine learning algorithms (Cardona et al., p. 1828, 2019). While all of the studies have a unique dataset, in general, the predictors encompassing previous academic performance, such as high school GPA, or standardized testing scores from the SAT or ACT, demographic, and socioeconomic information have been identified as being indicative of retention (Sandra C. Matz et al, p.2, 2023).

In their 2016 study, Alexandre M. Ohlbrecht, Christopher Romano, and Jeremy Tiegen approached predicting retention from a financial perspective. This perspective is reflected in three specific variables, the expected monetary contribution of family members, merit-based aid, and unmet need, which the authors define as the cost of attending the institution excluding the two aforementioned variables (Ohlbrecht et al., pg. 4, 2016). In conjunction, they used the variables of ethnicity, state residency (whether they were classified as in-state or out-of-state), campus residency status (whether they live on campus or commute), and standardized testing scores (Ohlbrecht et al., pg. 5, 2016). Using a dataset comprised of 4,523 first-time, full-time students from a public liberal arts college in New Jersey, and analyzing the students who did and did not return for their sophomore year, the study is broken down into an aggregate statistical summary, where using logistic regression, they found that higher standardized test scores (ACT

or SAT), being an on-campus resident, and whether or not a student has declared a major are strong predictors of retention. Regarding the financial concentration of the study, “as a family’s EFC increases, the likelihood of retention rises monotonically [and…] as the level of institutional financial assistance from the college to a student rises, the chance of retention for that student increases” (Ohlbrecht et al., p.10, 2016).

Similarly, Dursun Delen in his 2010 paper considered 16,066 freshmen, 2004-2008 from a public university located in the Mid-west of the United States (Dursun Delen, pg. 501, 2010). The foundation of Delen’s research is predicated on the data mining method entitled CRISP-DM, which involves six stages, “(1) understanding the domain and developing the goals for the study, (2) identifying, accessing, and understanding the relevant data sources, (3) pre-processing, cleaning and transforming the relevant data, (4) developing models using comparable analytical techniques, (5) evaluating and assessing validity and the utility of the models against each other and against the goals of the study, and (6) deploying the models for use in the decision making process” (Dursun Delen, p. 499, 2010). This process is depicted in Figure 1.1.

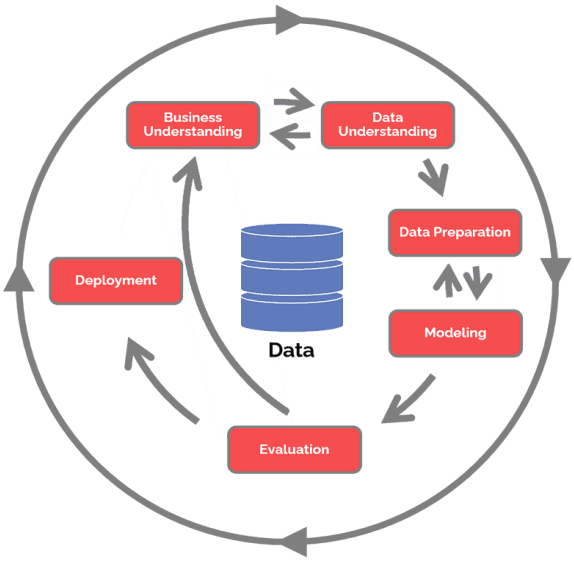


Figure 1.1 CRISP-DM Process

Using 39 variables from the dataset and 10-fold cross validation, Delen used artificial neural networks, decision trees, support vector machines, and three ensemble methods, bagging, busting, and information fusion, to predict retention (Dursun Delen, p. 501, 2010).

In his first analysis, where the class distribution for "Yes", a student retained, and "No", a student did not retain, was imbalanced, the support vector machine method had the highest prediction rate of 87.23%, followed by the decision tree method with a prediction rate of 87.3%, then the artificial network method with a prediction rate of 86.45%, and finally, the logistic regression method performed the worst with a prediction rate of 86.12% (Dursun Delen, p. 503, 2010). When looking at the students who were likely to drop-out after their first year, all four models performed poorly as they had a less than 50% accuracy (Dursun Delen, p. 503, 2010). Using the same methodology in his first analysis, he then performed the same analysis but where both of the classes for "Yes," and "No," were equal. Delen accomplished this by taking "all of the samples from the minority class (i.e., the "No" class herein) and randomly selected an equal number of samples from the majority class (i.e., the "Yes" class herein) and repeated this process for ten times to reduce bias of random sampling (Dursun Delen, p. 503, 2010). Although the prediction accuracies are different, model ranking-based performance was consistent with the results of the imbalanced class, that is, support vector machine had a prediction accuracy of 81.8%, decision tree had a prediction accuracy of 80.65%, artificial neural networks had a prediction accuracy of 79.85%, and logistic regression had an overall prediction accuracy of 74.26% (Dursun Delen, p. 503, 2010). Delen concluded that "regardless of the prediction model employed, the balanced data set (compared to unbalanced/original dataset) produced better prediction models for identifying the students who are likely to drop out of the college prior to their sophomore year" (Dursun Delen, p. 504, 2010)

While Alexandre M. Ohlbrecht, Christopher Romano, Jeremy Tiegen, and Dusun Delen used support vector machine (SVM) as one of their approaches when trying to predict retention based on freshman data, Tatiana A. Cardona, and Elizabeth A. Cudney, used it as their only approach in their 2019 paper. Their logic behind only implementing support vector machines was due to the small dataset they used which only contained 282 students from a Midwest community college, who were chemistry, biology, or engineering majors, and they had 9 associated variables (Tatiana A. Cardona et al., p. 1827-1831, 2019). Within their study, they classified retention as a student completing their degree within a three year-time span. This SVM was a type 2 classification, and used a radial basis function, in addition to using k-fold cross validation. Overall, "the model results showed a good performance with recall rates over 70% and testing rates over 78%" (Tatiana A. Cardona et al., p. 1931, 2019).

Similarly to Cardona and Cudney, Ruba Alkhasawneh and Rosalyn Hobson Hargraves focused on students in STEM disciplines when analyzing student retention and their motivation was "understanding the reasons behind the low enrollment and retention rates of Underrepresented Minority (URM) students (African Americans, Hispanic Americans, and Native Americans) in the disciplines of science, technology, engineering, and math" (Ruba Alkhasawneh et al., p.35, 2014). Within this 2014 study, Alkhasawneh and Hargraves define retention as students who stay enrolled in their respective STEM discipline, where the included majors are biology, chemistry, physics, science, forensic science, math, bioinformatics, environmental studies, computer and electrical engineering, biomedical engineering, mechanical engineering, chemical and life science engineering, from their first fall enrollment to their second (Ruba Alkhasawneh et al., p. 36, 2014). The primary group of students consisted of 1,966 full-time first-year STEM majors from 2007 through 2009, and a FeedForward backpropagation

network was used to model this. 10-fold cross validation was used to validate these neural network models and when feature selection was implemented the model's accuracy increased from 74% to 75% (Ruba Alkhasawneh et al., p. 40, 2014).

Sandra C. Matz approached retention by considering the extent to how accurately they can "predict whether a student is going to complete or discontinue their studies (in the future) by analyzing their demographic and socio-economic characteristics, their past and current academic performance, as well as their current embeddedness in the university system and culture" (Sandra C. Matz et al., p.2, 2023). Unlike previous researchers, Matz and her team used an app through their partnership with the educational software company READY Education. This app allows students to communicate with each other, even offering social media services, such as private messaging and groups, and this communication extends to any faculty member on campus, as well. For their study, this data was collected from 50,095 students from four separate institutions, and 462 features were extracted across all institutions. (Sandra C. Matz et al., p.2, 2023). In order to predict student retention, they used a linear classifier, elastic net, and a nonlinear classifier, random forest, however with both algorithms they used SMOTE in order to get more samples from the minority class, those who have not retained, from the data (Sandra C. Matz et al., p.2, 2023). Using AUC as their accuracy metric, typically, the random forest performed better at predicting retention, with an average AUC of 75%, than the elastic net, which had an average AUC of 70% (Sandra C. Matz et al., 2023, p.2). Finally, their results were consistent with previous research in which they found that student academic performance was an important predictor across all of their models and that since many of the engagement metrics on the app "are related to social activities or network features [it supports] the notion that a student's social

connections and support play a critical role in student retention" (Sandra C. Matz et al., 2023, p.10).

In March 2023, Dursun Delen, Behrooz Davazdahemami, and Elham Rasouli Dezfouli analyzed student attrition, through first conducting a comprehensive meta-analysis of eleven studies spanning from 2009 through 2023, which is shown in Table 1.1.

Table 1.1 Dursun et al. Machine Learning Algorithms to Predict Student Retention

Table 1 A summary of factors and approaches used in the literature for predicting attrition

Study	Subject	Approach	Factors Used for Prediction			
			Demographics	Educational	Financial	Socio-economic
(Delen, 2011)	Undergraduate Freshman Students	ANN	X	X	X	
(Berens et al., 2018)	Graduate students (German Universities)	AdaBoost	X	X		
(Thammasiri et al., 2014)	Undergraduate Freshman Students	SVM	X	X	X	
(Delen et al., 2020)	Undergraduate Freshman Students	Bayesian Belief Networks	X	X	X	
(Lin et al., 2009)	Engineering college students	ANN	X	X		
(Oztekin, 2016)	Undergraduate college students	SVM	X	X	X	
(Dissanayake et al., 2016)	Undergraduate college students	Random Forest	X	X		
(Fernández-García et al., 2021)	Undergraduate Freshman Students	Gradient Boosting Random Forest SVM Ensemble	X	X		
(Cannistrà et al., 2021)	Undergraduate Sophomore Students	GLM DT Random Forest	X	X		
(Baranyi et al., 2020)	Undergraduate Freshman Students	TabNet XGBoost Deep ANN	X	X		
Current study	Undergraduate Freshman Students	Deep Neural Networks	X	X	X	X

When comparing the various machine learning methodologies researchers used to predict student attrition, Delen and his fellow researchers determined that "the data used in those studies usually lack any features from one or more critical aspects thar are discussed in theoretical studies as the main determinants of the attrition decision, namely demographics, educational, financial, and socio-economic factors" (Dursun Delen et al., p.3, 2023). Within their study they addressed these gaps by using data of 39,470 freshmen enrolled in a mid-western United States university, where they defined retention as enrollment from fall semester of freshman year to the fall semester of their sophomore year (Dursun Delen et al., p.4, 2023). In conjunction to this data, they obtained the US annual GDP data per state from the Bureau of Economic Analysis website, obtained the average state-level per capitate income from the US Census Bureau, and

finally, obtained the annual state unemployment rate from the US Bureau of labor statistics, and mapped this information to the student based on their class status, freshman, and state of residence (Dursun Delen et al., p.4, 2023).

Once the data was compiled, they used a dense multi-layer perceptron deep neural network to predict retention based on the above features. The network that performed the best had an "overall accuracy of 88.4% in classifying students' attrition/retention status in their sophomore year" and looking specifically at the class distribution, the model had a 77.4% accuracy for students who dropped out, and 91.1% accuracy for students who enrolled (Dursun Delen et al., p.11, 2023). In order to determine which features were the most significant, Delen, Davazdahemami, and Dezfouli, used SHAP and found that "3 out of the top 5 factors all have to do with students' success in passing as many credits as possible with decent grades (GPA)" (Dursun Delen et al., p.13, 2023).

Chapter 2: Methodology

This chapter details the acquisition, processing, and cleaning of the various Ramapo College EOF data spreadsheets in order to learn more about the EOF student population and predict first-year retention. A summary of the techniques used, including each of the seven machine learning algorithms, is provided within this section.

Section 2.1: The Data Used for Exploratory Data Analysis

Twenty-one distinct Excel spreadsheets pertaining to semester grades for EOF students spanning from Fall 2013 to Spring 2023, were kindly provided by Dr. Nicole Videla, who is the senior director of the EOF program and Student Success. Each spreadsheet represented the academic report cards of the students for that semester, and consisted of the following 15 features, *term code, last name, first name, used first number, R number, email, subject code, course number, course section, course CRN, credit hours, grade, term attempted hours, term earned hours,* and *term GPA.* In an effort to maintain student privacy, the features *last name, first name, used first number, R number,* and *email* were dropped. A new feature entitled *Course Title* was appended to each dataframe, which was a result of concatenating the *subject code* and *course number.* Table 2.1.1 gives a list of the final features used within this study and their associated definition.

Table 2.1.1 EOF Report Cart Data Features

Features	Definition
Term Code	The year and semester
Subj Code	Subject area for the class
Crse Numb	Course number
Crse Sect	Course section
Crse CRN	Course registration number
Credit Hours	Credit hours
Grade	Letter grade recieved
Term Ahrs	Term attempted hours
Term Ehrs	Term earned horus
Term GPA	GPA for the semester
Course title	Subject code and course number for the class

Minimal data cleaning and preprocessing were required for this data. Each dataframe representing EOF student grades for each semester from Fall 2013 to Spring 2023 was then concatenated into one comprehensive dataframe, which is referred to as the EOF student grades dataframe, which contains 25,721 observations for the 11 features in Table 2.1.1. The first few observations from the EOF student grades dataframe, which shows the semester grades and courses for one student, are shown in Table 2.1.2.

Table 2.1.2 First 5 EOF Student Grades Observations

Term Code	Subj Code	Crse Numb	Crse Sect	Crse CRN	Credit Hours	Grade	Term Ahrs	Term Ehrs	Term GPA	Course Title
201340	ANTH	102	1	40959	4.0	B+	18.0	18.0	2.922	ANTH 102
201340	ANTH	220	1	40960	4.0	A-	18.0	18.0	2.922	ANTH 220
201340	EXSS	120	2	40637	2.0	A-	18.0	18.0	2.922	EXSS 120
201340	MATH	108	7	40210	4.0	C	18.0	18.0	2.922	MATH 108
201340	SOSC	235	6	40726	4.0	C+	18.0	18.0	2.922	SOSC 235

Section 2.2: EOF Retention Data

In order to build the dataset that would be used to predict EOF student first-year retention, twenty-one distinct excel sheet EOF rosters were shared with me by Dr. Videla for each semester between Fall 2013 – Spring 2023. For this analysis I excluded winter and summer sessions. Before any cleaning or preprocessing occurred, each roster had 32 features which pertained to student demographics, academic performance metrics, campus residency status, and academic advisors. In order to use any machine learning algorithms, the target variable of retention, whether a student retained or not after a year, had to be established.

For this analysis, retention is defined as a student being enrolled a year from their starting semester, for example, fall semester of their freshmen year to the fall semester of their sophomore year, or the spring semester of their freshman year to the spring semester of their sophomore year. While the majority of students will fall into this retention category, the definition also extends to students who transfer (i.e., their first year at Ramapo is as a sophomore, or junior). Based on this rationale, 18 dataframes were created to reflect a respective year of enrollment, Fall 2013- Fall 2014, Spring 2014- Spring 2015, …, Fall 2021- Fall 2022, Spring 2022- Spring 2023. Each dataframe was created by taking two rosters representing two distinct semesters, say Fall 2013 and Fall 2014. The students who had their entry term as Fall 2013 were then compared to the students' roster, of Fall 2014. If the student with entry term Fall 2013 roster was also in the Fall 2014 roster, then in the new combined dataframe, reflecting the year Fall 2013- Fall 2014, the student had a *Y* to indicate they retained in the new *Retention* column. Otherwise, if the student with the entry term of Fall 2013 was not in the Fall 2014 dataframe, the student had an *N* in the *Retention* column. This process is depicted in Figure 2.2.1.

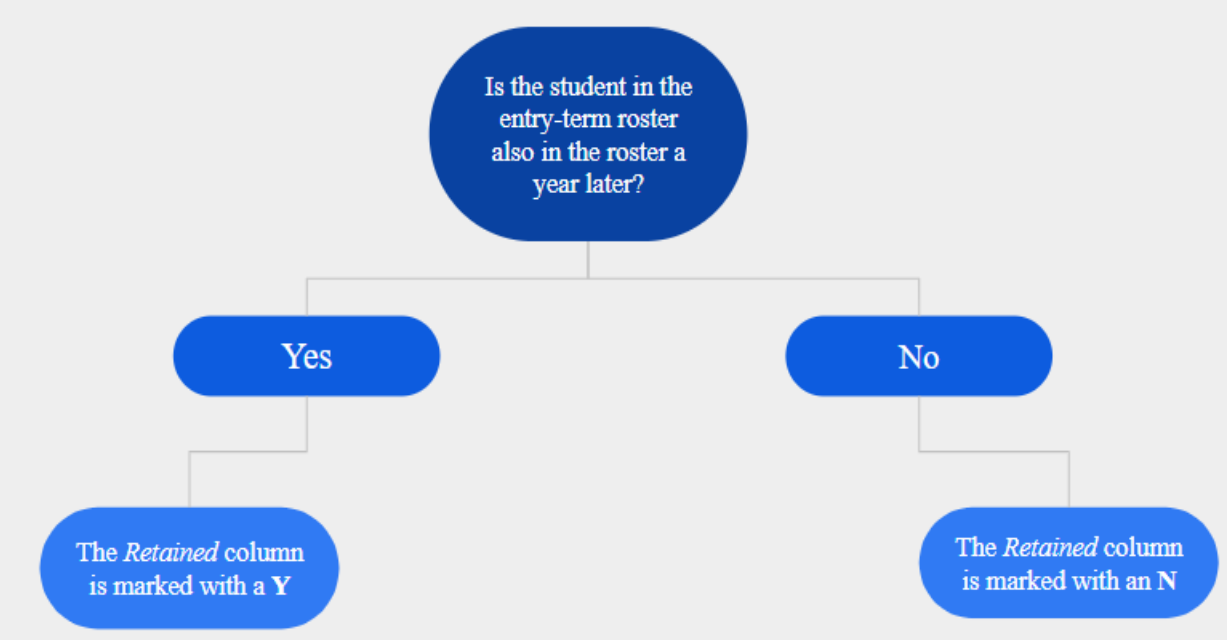


Figure 2.2.1 EOF Retention Definition

The same process was then repeated for the other 17 academic years, and then all of these dataframes reflecting enrollment for Fall 2013- Fall 2014, Spring 2014- Spring 2015, …, Fall 2021- Fall 2022, Spring 2022- Spring 2023, were then concatenated into a single dataframe, which is referred to as the EOF retention dataframe for the rest of this analysis.

The EOF retention dataframe was modified further, as several features were manipulated and dropped. The feature *Birth Date* was renamed to be *Age* and is an approximate estimate that was determined by extracting the year of the students' entry term and subtracting the year of their birth date. Since this is an evaluation of EOF students, I dropped academic advisors, but retained the EOF advisors. The following features were dropped, *Last Name, First Name, Used First Name, R Number, Email, Race2, Levl, Styp Desc, Conc 1, Advisor 1, Advisor 2, TRIO, First Gen,* and *Upward Bound,* as they either revealed identifying information about the student or had inconsistent data. The 18 features and target variables, along with their description, and datatype, which is used throughout the rest of this analysis are detailed in Table 2.2.1.

Table 2.2.1 EOF Retention Features

Feature	Description	Data Type
Gender	*M* for Male or *F* for Female	Categorical
New Ethn	Ethnicity of student, with the categories *Not Hispanic or Latino* or *Hispanic or Latino*	Categorical
Race1	Race of student with the categories,*Black or African American*, *White*, *Asian*, *American Indian or Alaskan Native*	Categorical
Class	*FR* for Freshman, *SO* for Sophomore, *JR* for Junior, *SR* for Senior	Cateogorical-ordinal
Styp Code	*N* for new first time student, *W* for continuing first time student, *T* for transfer, *U* for continuining transfer, *R* for readmit, a student who takes a leave of absence or is out for at least a year, *C* for continuining readmits, *B* for second bachelor's, and *M* for non-matriculated	Categorical
School	School within the college, *SB* for Anisfield School of Business, *SS* for School of Social Science and Human Services, *CA* for School of Contemporary Arts, *TS* for School of Theoretical and Applied Science, and *HG* for School of Humanities and Global Studies	Categorical
Majr 1	Student's major as catalogoued by the college	Categorical
Term Reg Hrs	Registered credit hours for the semester	Numeric
Term Ahrs	Attempted credit hours for the semester	Numeric
Term Ehrs	Earned credit hours for the semester	Numeric
Term GPA	GPA for the semester	Numeric
Cum Ahrs	Cumulative attempted credit hours	Numeric
Cum Ehrs	Cumulative earned credit hours	Numeric
Cum GPA	Cumulative GPA	Numeric
Matric Term	First term that the student started at Ramapo	Numeric
Resd?	Whether a student is a resident on campus, *Y*, or a commuter, *N*	Categorical
Retained	Whether a student has been enrolled for a full year since starting at Ramapo, *Y*, or they have not have not, *N*	Categorical
EOF Advisor	The student's EOF advisor during that that semester	Categorical
Age	Apprxoimate age of the student based on their birth year and entry term	Numeric

For the students who did not have a listed EOF academic advisor, in addition to where there were missing values for other features, these students were dropped from the EOF retention dataframe resulting in 515 observations. The first few observations of the final, cleaned version of the EOF retention dataframe with 515 rows and 19 features, is shown in Table 2.2.2.

Table 2.2.2 First 5 Observations of Cleaned EOF Retention Dataframe

	Gender	New Ethn	Race1	Class	Styp Code	School	Majr 1	Term Reg Hrs	Term Ahrs	Term Ehrs	Term GPA	Cum Ahrs	Cum Ehrs	Cum GPA	Matric Term	Resd?	Retained	EOF Advisor	Age
0	F	Not Hispanic or Latino	Black or African American	FR	N	TS	BIOL	18	18	14	2.556	26	22	2.892	201340	Y	Y	EOF Erika Vega	18
1	F	Hispanic or Latino	White	FR	T	AI	LITR	16	16	42	3.567	16	42	3.567	201340	Y	Y	EOF Marita Esposito	19
2	F	Not Hispanic or Latino	Black or African American	FR	N	TS	BIOL	18	18	18	2.844	26	22	3.092	201340	Y	Y	EOF Erika Vega	19
3	F	Not Hispanic or Latino	Black or African American	FR	N	TS	BIOL	15	15	24	3.571	23	32	2.927	201340	Y	Y	EOF Erika Vega	18
4	F	Not Hispanic or Latino	White	FR	N	SB	UNDC	18	18	14	2.957	26	18	2.973	201340	Y	Y	EOF Nicole Videla	18
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
510	M	Hispanic or Latino	White	FR	N	SS	PSYC	16	16	12	1.500	20	12	1.860	202140	N	N	EOF Tushawn Jernigan	19
511	M	Not Hispanic or Latino	Asian	FR	N	TS	NURG	17	17	25	3.929	21	29	3.943	202140	N	Y	EOF Natalie Quiñones	18
512	F	Not Hispanic or Latino	Asian	FR	N	TS	NURG	17	17	29	3.694	21	33	3.752	202140	Y	Y	EOF Natalie Quiñones	18
513	F	Not Hispanic or Latino	Asian	FR	T	SB	INFM	16	16	28	0.000	16	28	0.000	202140	Y	Y	EOF Deirdre Bright Foreman	21

The EOF retention dataframe was then split into two separate datafames representing EOF student retention pre-covid, and EOF student retention post-covid, in order to have a more robust analysis. For this study, pre-covid retention is defined as students with starting semesters of Fall 2013 through Spring 2019, and post-covid retention is defined as students with starting semesters from Spring 2020 through Spring 2023. Throughout the rest of this paper, these dataframes are respectively referred to as the pre-covid EOF retention dataframe, and post-covid EOF retention dataframe. Both dataframes have 19 columns (18 features and 1 target), however, the pre-covid EOF retention dataframe has 406 observations whereas the post-covid EOF retention dataframe has 114 observations. Since the first few observations of the pre-covid EOF retention dataframe are the same as the first few observations of the EOF retention dataframe, the first three observations of the post-covid EOF retention dataframe are shown in Table 2.2.3.

Table 2.2.3 First 5 Observations of Post-Covid EOF Retention Dataframe

	Gender	New Ethn	Race1	Class	Styp Code	School	Majr 1	Term Reg Hrs	Term Ahrs	Term Ehrs	Term GPA	Cum Ahrs	Cum Ehrs	Cum GPA	Matric Term	Resd?	Retained	EOF Advisor	Age
0	F	Hispanic or Latino	White	JR	T	SS	SWRK	12	12	77	3.333	12	77	3.333	202020	N	N	EOF Andre Turner	25
1	F	Hispanic or Latino	White	JR	T	SB	ACCT	16	16	81	3.175	16	81	3.175	202020	N	Y	EOF Tushawn Jernigan	23
2	F	Hispanic or Latino	White	SO	T	CA	MUSI	12	12	46	3.350	12	46	3.350	202020	Y	Y	EOF Andre Turner	20
3	M	Not Hispanic or Latino	White	FR	N	TS	BIOL	18	18	18	3.011	22	22	3.191	202040	N	Y	EOF Tushawn Jernigan	18
4	F	Hispanic or Latino	White	FR	N	CA	COMM	16	16	12	4.000	20	16	4.000	202040	Y	Y	EOF Andre Turner	19
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
109	M	Hispanic or Latino	White	FR	N	SS	PSYC	16	16	12	1.500	20	12	1.860	202140	N	N	EOF Tushawn Jernigan	19
110	M	Not Hispanic or Latino	Asian	FR	N	TS	NURG	17	17	25	3.929	21	29	3.943	202140	N	Y	EOF Natalie Quiñones	18
111	F	Not Hispanic or Latino	Asian	FR	N	TS	NURG	17	17	29	3.694	21	33	3.752	202140	Y	Y	EOF Natalie Quiñones	18
112	F	Not Hispanic or Latino	Asian	FR	T	SB	INFM	16	16	28	0.000	16	28	0.000	202140	Y	Y	EOF Deirdre Bright Foreman	21

There are six machine learning algorithms that I implement throughout this analysis, logistic regression, decision tree, random forest, support vector machine, gradient boosting classifier, and ensemble learning (random forest, logistic regression, and support vector machine simultaneously), to predict retention for EOF students. The application of each algorithm is divided into three sections, predicting retention for all EOF students, predicting retention for all EOF students pre-covid, and predicting retention for all EOF students post-covid, and use the three dataframes as detailed above EOF retention, pre-covid EOF retention, and post-covid EOF retention, respectively. Although it is detailed extensively later in this study, there is a substantial class imbalance within the entire EOF population of students who have retained, 452, versus students who have not retained, 60. To address this class imbalance, the Synthetic Minority Oversampling Technique, SMOTE, are used, which creates new observations of the minority

class. Feature selection will also be a consideration for each algorithm, and are determined using Shapley Additive Explanation, SHAP, which uses game theory to assign importance scores for predictors which are determined by their contribution to predicting performance, which for this thesis is retention (SHAP documentation).

For each machine learning algorithm implemented, there are 12 iterations, as four analyses will occur in each section – all EOF students, all EOF students pre-covid, and all EOF students post-covid. Within each section, a model is fit using all of the respective EOF retention data, using all of the retention data, and applying smote, using all of the retention data and using SHAP for feature selection, and finally using all of the retention data and applying both smote and feature selection. The visualization of this process, and how 12 models are fit for a given machine learning algorithm is shown in Figure 2.2.2.

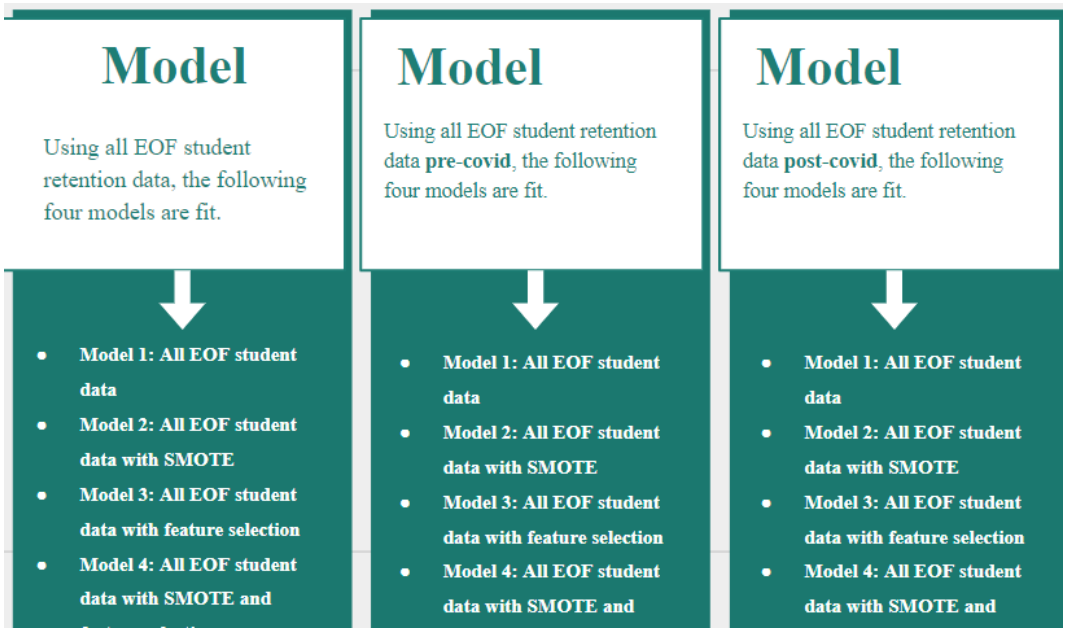


Figure 2.2.2 Predicting EOF First-Year Retention Process

In the following sections, I provide a summary of the techniques and machine learning algorithms that are utilized throughout this study.

Section 2.3 Synthetic Minority Oversampling Technique (SMOTE)

The synthetic minority oversampling technique, (SMOTE), is a preprocessing method that was developed to combat the overfitting of models. This typically occurs as a result of the standard random sampling approach (Fernandez et al., p. 864, 2018) when the outcome groups are imbalanced. As in my example, the 455 students retained is much larger than the 60 that were not. This approach creates new instances of the minority class (not retained) to have a more balanced comparison of outcomes. This method was first proposed in 2002 and is viewed as an influential preprocessing and sampling method in machine learning (Fernandez et al., p. 864, 2018).

Section 2.4 Shapley Additive Explanation (SHAP)

SHapley Additive exPlanations (SHAP), relies on concepts from coalitional game theory to "provide an explanation for a machine learning model's prediction by computing the contribution of each feature to the prediction" (Soufiane Fadel). Game theory studies optimum decision-making by competing agents, also known as players, within a game. Cooperative game theory can be derived from this framework, which proposes that these players drive decision-making and generate cooperative conduct. Rather than being a game between individual players, the perspective now shifts, and it is viewed as a competition between an alliance of players. In order to measure this, 'the shapely value is defined as the marginal contribution of variable value to prediction across all conceivable 'coalitions' or subsets of features" (Soufiane Fadel).

All Shapley values satisfy the following properties – efficiency, symmetry, dummy, and linearity, which, when viewed and calculated together within this context, are assumed to represent a fair weight. The implementation of the Shapley value within an analysis "is often

preferable since it is based on a solid theory, fairly divides the effects, and provides a complete explanation" (Soufiane Fadel).

Section 2.5 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction algorithm that not only reduces training time, but facilitates data visualization (Aurelien Geron, p.213, 2020). Simplistically, PCA involves two distinct steps, "first it identifies the hyperplane that lies closest to the data, and then it projects the data onto it" (Aurelien Geron, p. 219, 2020). In decreasing order, PCA determines the axes that contain the highest variance within the training set, and the i^{th} axis is referred to as the i^{th} principal component (Aurelien Geron, p. 220, 2020). These principal components of the training set are determined by using "the standard matrix factorization technique called *Singular Value Decomposition* (SVD) that can decompose the training set matrix X into the matrix multiplication of three matrices" (Aurelien Geron, p. 221, 2020). In order to determine the correct hyperplane to ensure the maximum amount of variance is preserved, the *explained_variance_ratio* method is used from Scikit. For this analysis, the number of principal components is reduced to 2 dimensions to visualize the data, and use k-means clustering (Aurelien Geron, p. 223, 2020).

Section 2.6 K-Means Clustering

K-means clustering is an unsupervised learning technique proposed by Stuart Lloyd in 1957 that takes data, referred to as an instance, and assigns it to a group of similar instances, which is referred to as a cluster (Aurelien Geron, p.236, 2020). K-means clustering can only be applied to numerical data and adheres to an iterating algorithm which begins with randomly assigning centroids. With each iteration, the instances are assigned to a cluster, and then the cluster centroid is computed. If an instance is closer to the centroid of a different cluster, then it

is assigned to that cluster and the centroids computed again. The cluster centroids will eventually stabilize and those are the K-means clusters.

In order to determine the optimal number of clusters, k, the silhouette score is implemented. This score is the mean of the silhouette coefficient over all of the data instances, where "an instance's silhouette coefficient is equal to $\frac{(b-a)}{max(a,b)}$ where a is the mean distance to the other instances in the same cluster (i.e., the mean intra-cluster distance) and b is the mean nearest-cluster distance (i.e., the mean distance to the instances of the next closest cluster, defined as the one that minimizes b, excluding the instance's own cluster) (Aurelien Geron, p.246-247, 2020). This value for the silhouette score is usually within the range of -1 to 1.

Section 2.7 Logistic Regression

Logistic Regression is a supervised learning binary classifier that estimates the probability that an observation belongs to a class using a specific threshold. Typically, this value is 50%, and if the estimated probability is over 50% then the model predicts that this observation belongs to a different class than if the estimated probability is under 50 (Aurelien Geron, p. 142, 2020). Before any class predictions can be estimated, all categorical variables must be encoded, which for this analysis was accomplished using the *get_dummies* method. For example, the target variable, *Retained,* the *Y* responses became 1 and the *N* responses became 0. After this method was applied, the number of columns increased from 19 (18 features and 1 target variable) to 80 (79 encoded features and encoded target variable) for all three dataframes – EOF retention, pre-covid EOF retention, and post-covid EOF retention.

Logistic Regression works similarly to Linear Regression, as the "model computes a weighted sum of the input features (plus a bias term), but instead of outputting the result directly like the Linear Regression model does, it outputs the logistic of this result," (Aurelien Geron, p.

143, 2020). The logistic is a sigmoid function that produces a value between 0 and 1, meaning a probability of being retained (Aurelien Geron, p. 143)

Section 2.8 Decision Tree and Random Forest

A decision tree is a supervised learning algorithm that is the basis of the random forest algorithm and can be used for both classification and regression tasks (Aurelien Geron, p. 175, 2020). However, where a decision tree is easier to interpret and is a white box model, a random forest is considered more difficult to interpret and is considered a black box model. In order to predict EOF student first-year retention, the decision tree classifier is used, and similarly to the logistic regression algorithm, all categorical variables must be encoded, which was accomplished using the *get_dummies* method. After this method was applied, the number of columns increased from 19 (18 features and 1 target variable) to 80 (79 encoded features and encoded target variable) for all three dataframes – EOF retention, pre-covid EOF retention, and post-covid EOF retention.

The model is most easily explained through the associated visualization of the decision tree. The classification of an instance is determined by starting at the root node, and following the conditions of it and subsequent nodes, until a prediction result is made at a leaf node. A node that does not have an associated branching is referred to as a leaf node. Each node is comprised of a gini, samples, value, and class. The sample of a node explains how many training instances are applicable to this condition, whereas the value provides the distribution of the training instances applicability to each of the classes, which in this case would be 0, for not retained, and 1, to indicate retention. The "gini attribute measures its *impurity*: a node is 'pure' (gini-0) if all training instances it applies to belong to the same class", and its equation is as follows, $G_i = 1 - \sum_{k=1}^{n} P_{i,k}^2$, where "$P_{i,k}$ is the ratio of class k

Instances among the training instances in the i^{th} node" (Aurelien Geron, p. 177, 2020).

Section 2.9 Random Forest

A Random Forest is an ensemble supervised machine learning algorithm comprised of decision trees that can perform both classification and regression tasks (Aurelien Geron, p. 197, 2020). A decision tree is easier to interpret and is a white box model, while a random forest is considered more difficult to interpret and is considered a black box model. In order to predict EOF student first-year retention, the random forest classifier is used, and all categorical variables must be encoded using the *get_dummies* method. In general, the "algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node, it searches for the best feature among a random subset of features" (Aurelien Geron, p. 197, 2020).

Section 2.10 Support Vector Machine

Support Vector Machine is a powerful supervised machine learning method that can perform linear classification, nonlinear classification, regression, and outlier detection (Aurelien Geron, p. 153, 2020). The type of SVM method used is most easily determined by examining representative model plots with associated decision boundary lines. If two classes can be distinctly separated by a straight line, they can be considered linearly separable and it is advisable for linear SVM classification to be implemented, which is the case for this study (Aurelien Geron, p. 153, 2020). Similar to the preprocessing for other methods all categorical variables must be encoded, using the *get_dummies* method.

If the decision boundaries are viewed as a street, the addition of "more training instances "off the street" will not affect the decision boundary at all: it is fully determined (or "supported")

by the instances located on the edge of the street", where these are instances are referred to a support vectors (Aurelien Geron, p. 155, 2020).

Section 2.11 Gradient Boosting Classifier

Gradient Boosting is a supervised ensemble machine learning algorithm that "works by sequentially adding predictors to an ensemble, each one correcting its predecessor" (Aurelien Geron, p. 203, 2020). Unlike other methods, such as AdaBoost, where the instance weights are changed at every iteration, the residual errors made by the previous predictor are fit to the new predictor (Aurelien Geron, p. 203, 2020). Similar to the other methods mentioned in this section, gradient boosting can perform both regression and classification tests. For the purpose of this study, predicting EOF student first-year retention, the Gradient Boosting Classifier are used. Similar to the preprocessing for previous methods, all categorical variables must be encoded, using the *get_dummies* method. After this method was applied, the number of columns increased from 19 (18 features and 1 target variable) to 80 (79 encoded features and encoded target variable) for all three dataframes – EOF retention, pre-covid EOF retention, and post-covid EOF retention.

Section 2.12 Ensemble Learning (Logistic Regression, Random Forest, Support Vector Machine)

An ensemble learning algorithm is comprised of a group of predictors (which could be classifiers or regressors), which when aggregated, typically produce a better predictor (Aurelien Geron, p. 189, 2020). For this analysis, the ensemble learning method consists of logistic regression, random forest classifier, and support vector machine, algorithms. In order to predict the class an observation belongs to, retained, or not retained, a hard voting classifier are used which aggregates the predictions from each respective algorithm, and then predicts the class that

received the most votes (Aurelien Geron, p. 190, 2020). The preprocessing for each of these algorithms was the same as described in their respective sections, using the *get_dummies* method.

The hard voting classifier "often achieves a higher accuracy than the best classifier in the ensemble. In fact, even if each classifier is a weak learner (meaning it does only slightly better than random guessing), the ensemble can still be a strong learner (achieving high accuracy), provided there are a sufficient number of weak learners and they are sufficiently diverse" (Aurelien Geron, p. 190, 2020).

Section 2.13 K-fold Cross-Validation

The purpose of this work is to find a model that is able to perform well on data it has not been trained on (i.e., new first-year students). This method of k-fold cross validation is used to simulate this work, and within this analysis, $k = 10$. With 10-fold validation, the training dataset is split into 10-folds, and predictions are made using on each fold using a model that was trained on the remaining 9 folds (Aurelin Gero, p. 90, 2020). After all predictions are made on each fold, the summarized data provides an evaluation score for that model. This should be representative of how the model will perform on untested data.

Chapter 3: The EOF Student Population

All information within this section is performed using the – EOF retention, pre-covid EOF retention, and post-covid EOF retention data, which were discussed in the methodology section. All bar graphs are sorted in decreasing order based on the count of EOF students who fall into the *Not Retained*, category for the particular feature being examined, since this result is more interesting and informative to the EOF department faculty.

Section 3.1 Examining EOF Student Retention and Age

Using the EOF retention dataframe, the distribution of students who have and have not retained is shown in the following bar graph. As shown in Figure 3.1.1, the ages with the highest number of students who have not retained are age 18, with a count of 37 students not retaining, and age 19, with a count of 18 students not retaining. However, the ages that have the highest number of students retaining are also ages 18, with a count of 291 students, and age 19, with a count of 112 students.

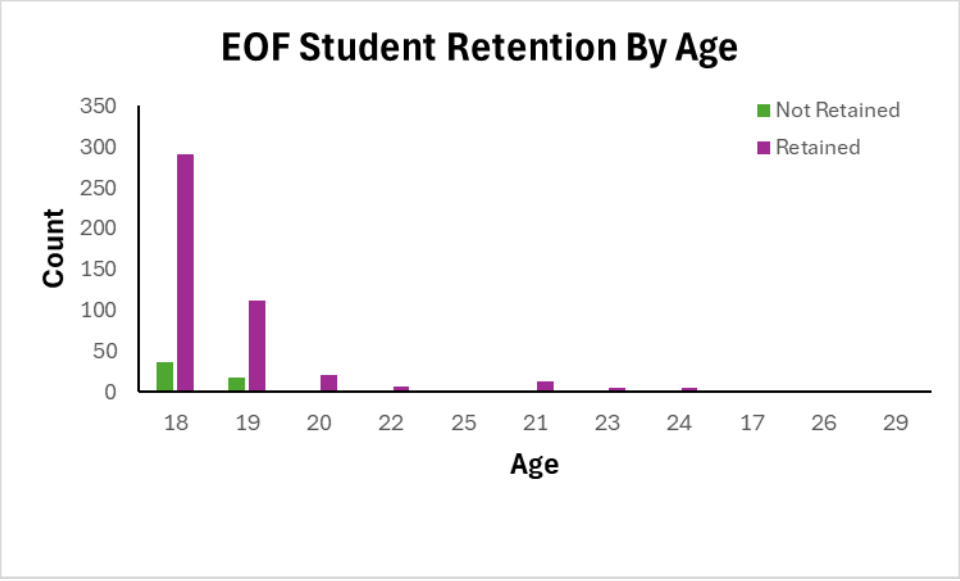


Figure 3.1.1 EOF Student Retention By Age

The results are consistent when the analysis is broken down into pre-covid and post-covid EOF students' retention, respectively. Ages 18 and 19 still have the highest counts of student retention and attrition. Looking at pre-covid EOF student attrition first, which is shown in Figure 3.1.2, for age eighteen 24 students did not retain, and for age nineteen, 9 students did not retain. Regarding retention, for age eighteen 234 students were retained for age nineteen, 90 students were retained.

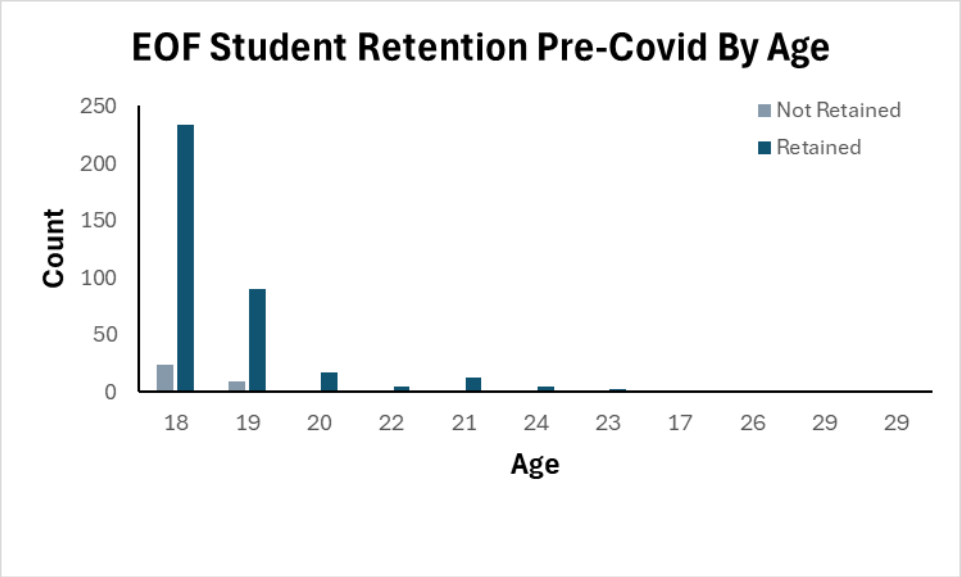


Figure 3.1.2 EOF Student Retention Pre-Covid By Age

Examining the post-covid EOF retention distribution by age, which is shown in Figure 3.1.3, for ages 18 and 19, 13 students and 9 students, respectively, did not retain, and 57 and 22 students, respectively, did retain.

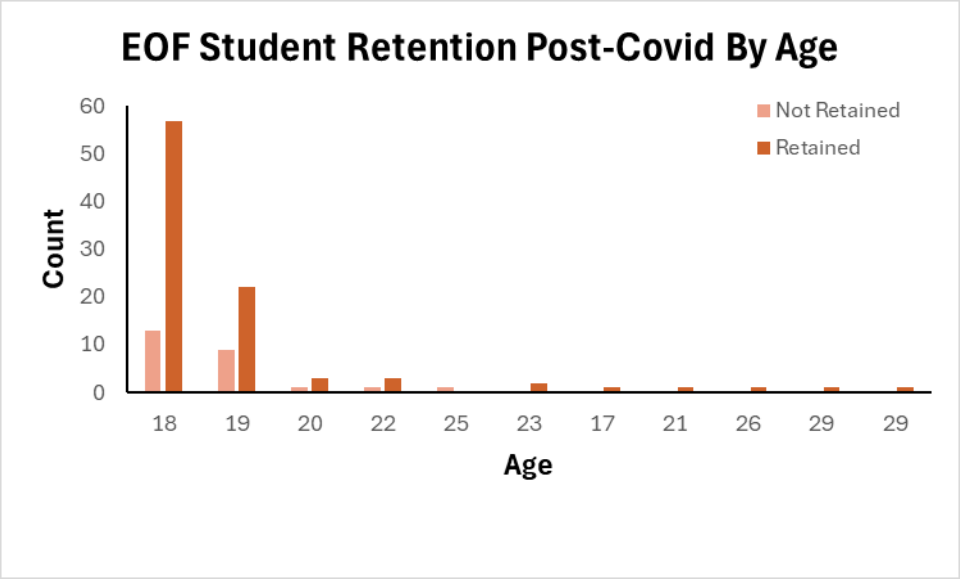


Figure 3.1.3 EOF Student Retention Post-Covid By Age

The consistency of the ages for the highest counts of retention and attrition, based on Figures 3.1.1, 3.1.2, and 3.1.3, suggests that at ages 18 and 19, EOF students are more susceptible to attrition.

Section 3.2 Examining EOF Student Retention and Major

Within this section, the relationship between EOF student retention and their declared major is examined. Figure 3.2.1 shows the count of students per declared major, who did and did not retain. As the figure demonstrates, the top four majors with the highest counts of EOF student attrition are undeclared (UNDC), psychology (PSYC), social work (SWRK), and biology (BIOL), with respective counts of 10, 8, 5 and 5, students.

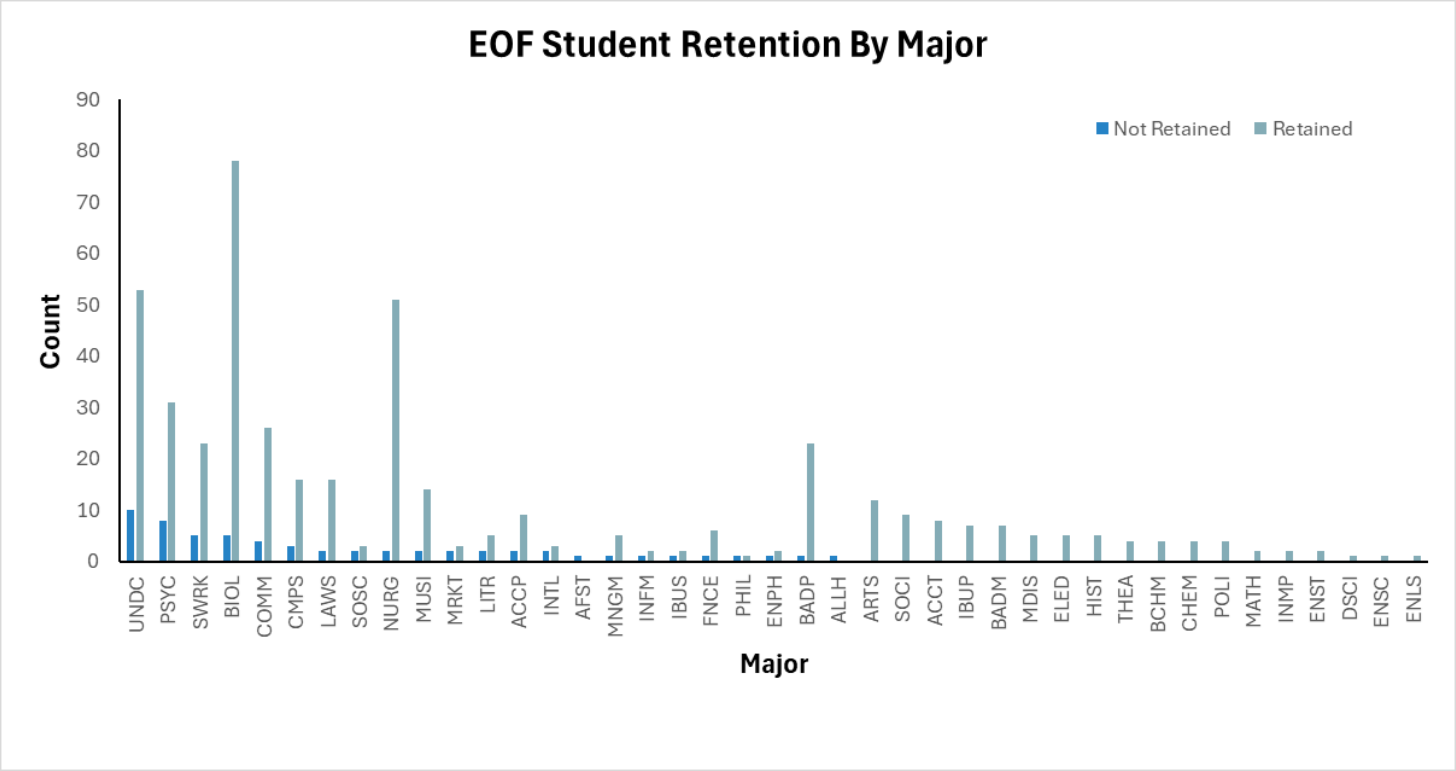


Figure 3.2.1 EOF Student Retention By Major

Looking at the distribution of EOF student retention by major pre-covid, which is shown in Figure 3.2.2, the results are similar to the distribution of EOF student retention by major for all students spanning the data as shown in Figure 3.2.1. While the top four majors with the highest attrition are the same, the order and associated counts are as follows, the undeclared major has 8 instances of students not retaining, the psychology major has 4 instances of students not retaining, and the biology and social work majors have 3 instances of students not retaining. This is shown in Figure 3.2.2.

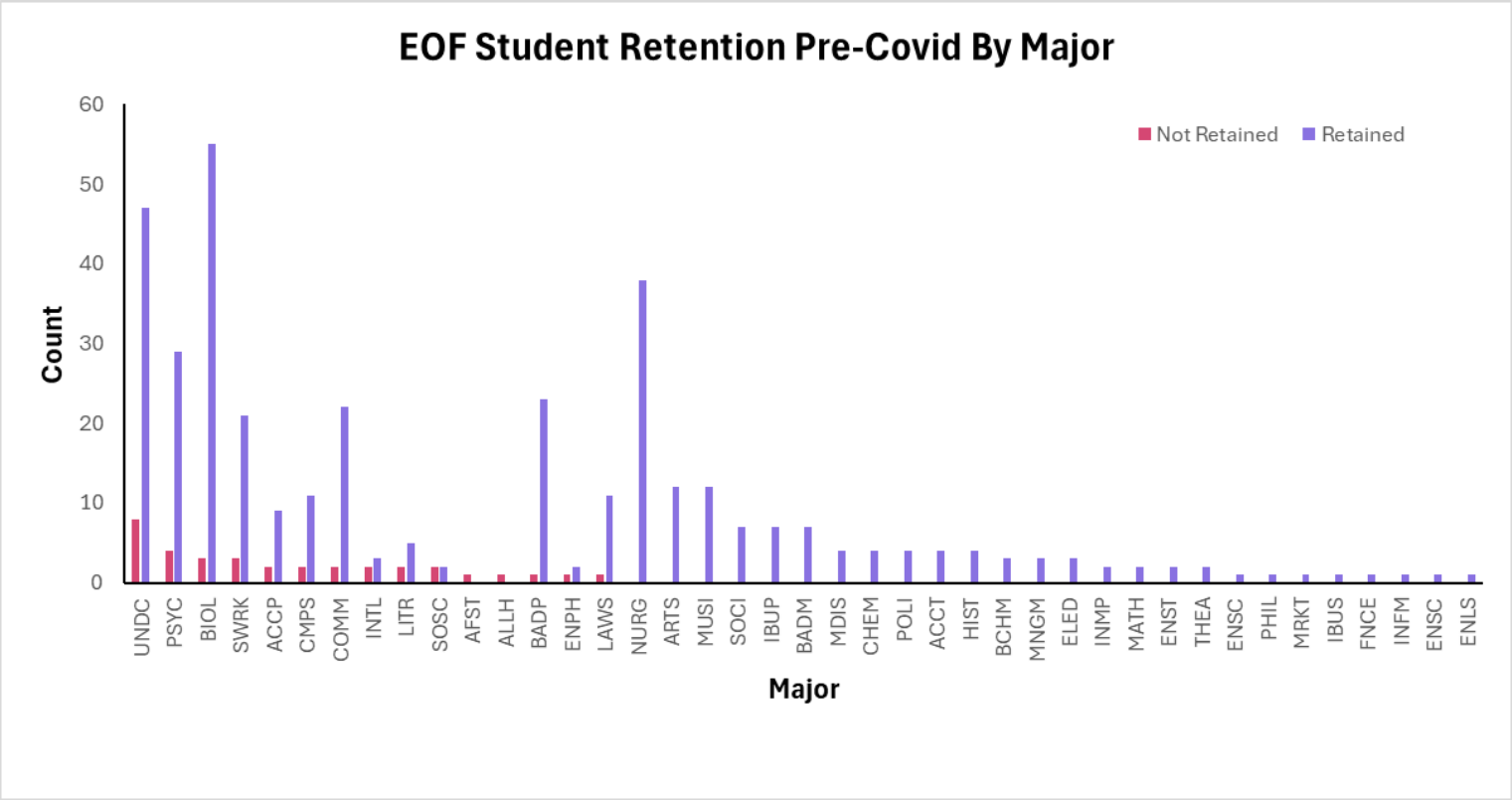


Figure 3.2.2 EOF Student Retention Pre-Covid By Major

The distribution of retention and attrition for EOF students post-covid, which is depicted in Figure 3.2.3, is different compared to distributions for all EOF students, and all EOF students pre-covid. Where the psychology and biology majors comprised the second half of majors with the highest attrition counts for all EOF students and all EOF students pre-covid, for all EOF students post-covid, psychology and biology were the majors with the highest counts of students, 4 and 2, respectively, of students who have not retained. Communication (COMM) and marketing (MRKT) majors are next with a count of 2 students who have not retained each.

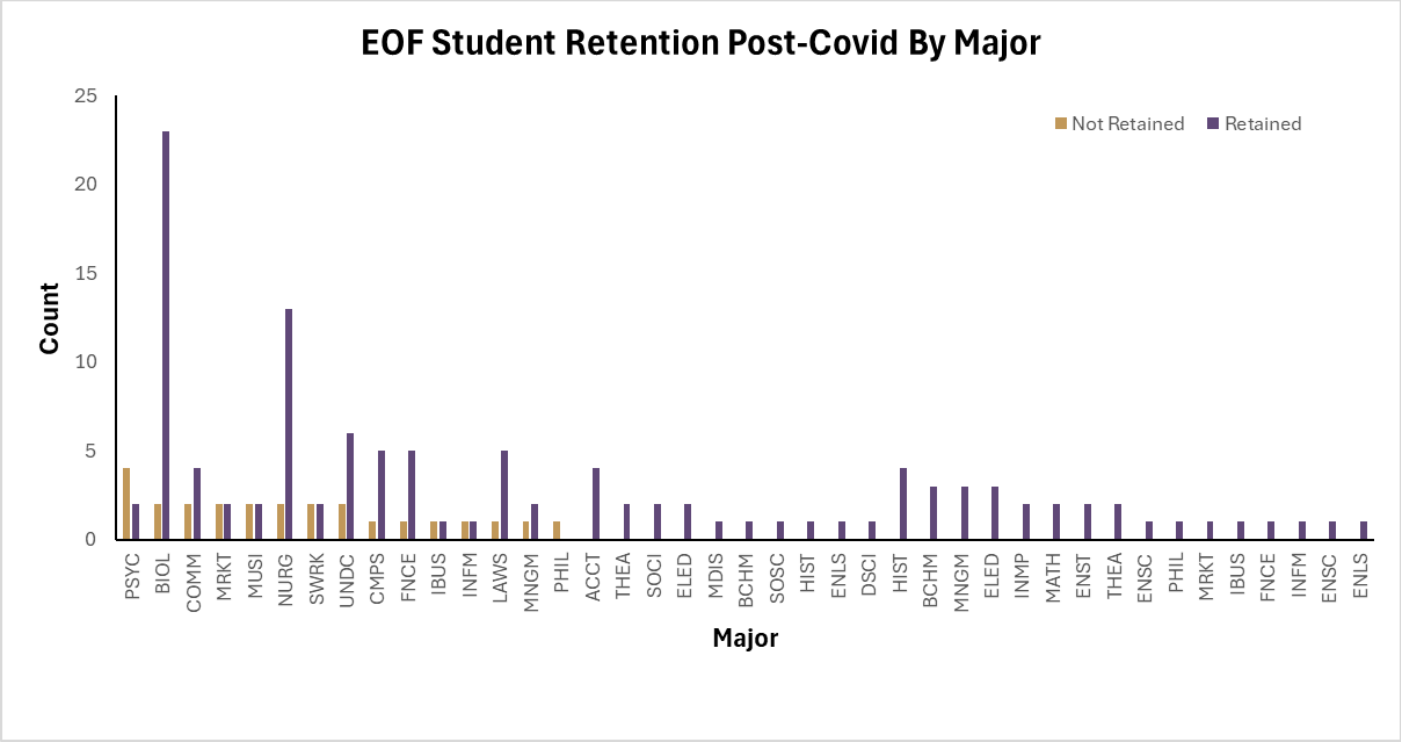


Figure 3.2.3 EOF Student Retention Post-Covid By Major

Based on Figures 3.2.1, 3.2.2, and 3.2.3, this suggests that students who major in psychology, biology, social work, communication, marketing or are undeclared, tend to have higher attrition rates.

Section 3.3 Examining EOF Student Retention and School

Within this section, the relationship between EOF student retention and the school a student is enrolled in (at Ramapo College), is explored. Figure 3.3.1 shows that the School of Social Science and Human Services (SS), the School of Theoretical and Applied Science (TS), and the Anisfield School of Business (SB) have the highest occurrences of student attrition, with counts of 20, 13, and 9 students not retaining, respectively.

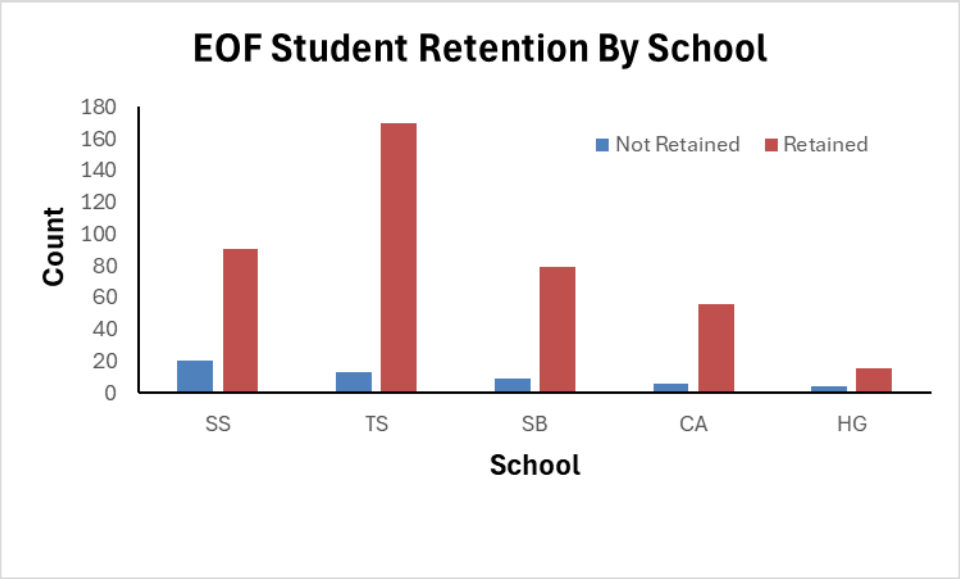


Figure 3.3.1 EOF Student Retention By School

When considering the distribution of attrition and retention for EOF students pre-covid and post-covid, for both analyses, the highest instance of attrition occurs within the School of Social Science and Human Services, with counts of 13 and 7 students not retaining, respectively. However, for pre-covid, the School of Theoretical and Applied Science follows this result, with 7 students not retaining, and then the School of Humanities and Global Studies, with 3 students not retaining. This is shown in Figure 3.3.2.

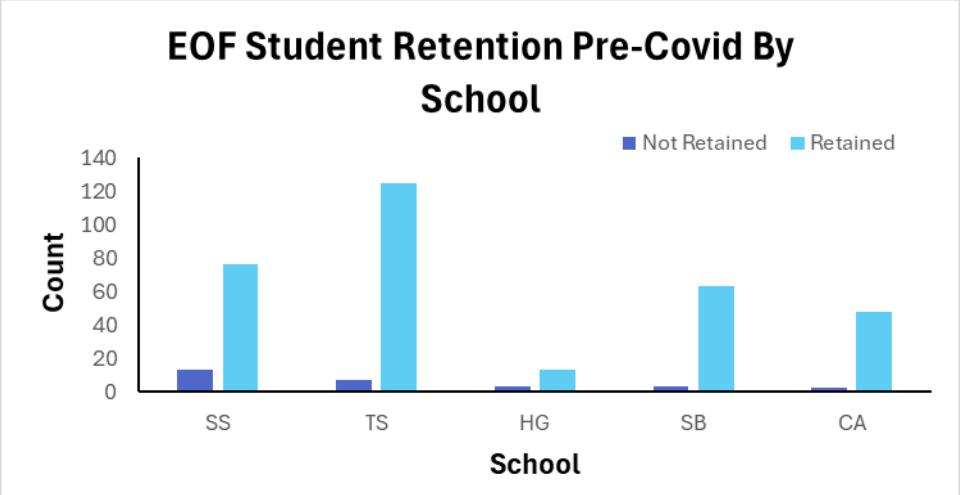


Figure 3.3.2 EOF Student Retention Pre-Covid By School

For EOF students post-covid, the top three schools with the highest instances of EOF student attrition are the same as for all EOF students. However, Anisfield School of Business then the School of Theoretical and Applied Science are the same, with 6 students not retaining each. This is shown in Figure 3.3.3.

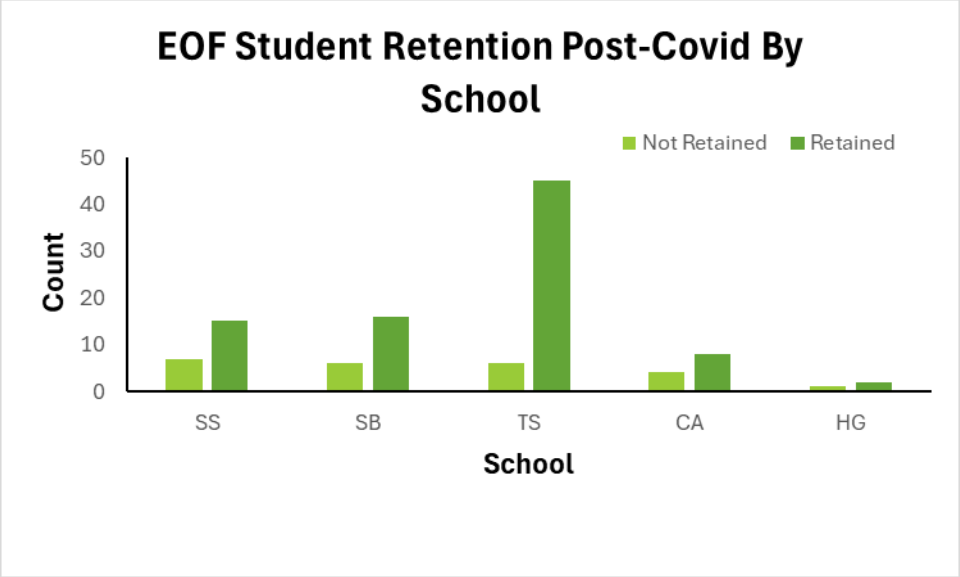


Figure 3.3.3 EOF Student Retention Post-Covid By School

The results within this section suggest, based on Figures 3.3.1, 3.3.2, and 3.3.3, that the school of Social Science and Human Services, the School of Theoretical and Applied Science, the Anisfield School of Business, and the School of Humanities and Global Studies, have the highest occurrences of student attrition.

Section 3.4 Examining Student Retention and Gender

The distribution of EOF student retention is explored within the context of gender. Within this study, gender is comprised of Male, *M*, and Female, *F*, as this was the format in which the data was provided. Figure 3.4.1 shows that females have a higher count of attrition, 38 instances, than males, with 22 occurrences. However, when looking at the total EOF female

population, 10.4% of female students did not retain, whereas for the total male EOF population 14.8% of the male students did not retain.

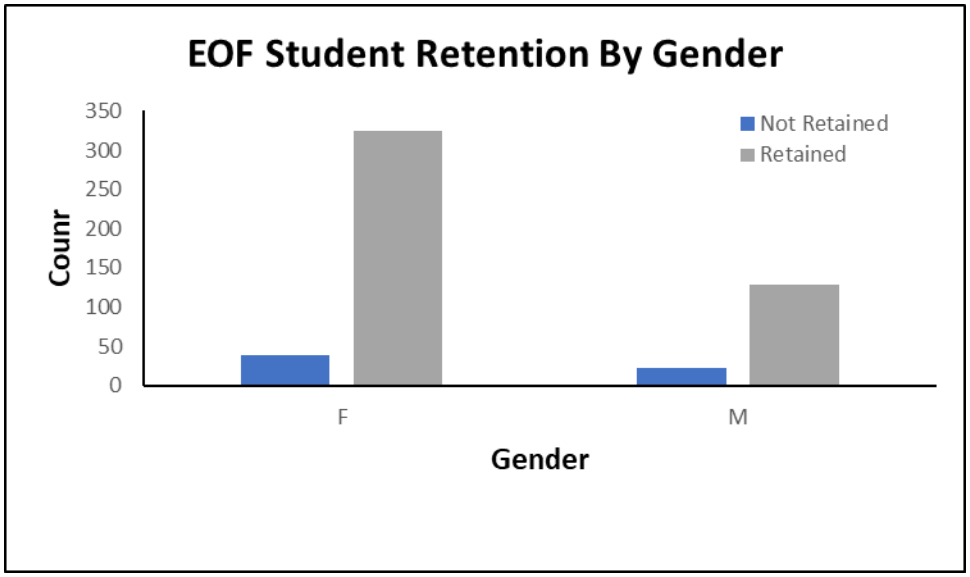


Figure 3.4.1 EOF Student Retention By Gender

This result of women having a higher count of attrition, than males is consistent for both the pre-covid analysis and post-covid analysis. Figure 3.4.2 shows that 22 females did not retain, or 7.8% of the EOF female population pre-covid did not retain, and 13 males did not retain, or 11% of the EOF male population post-covid did not retain.

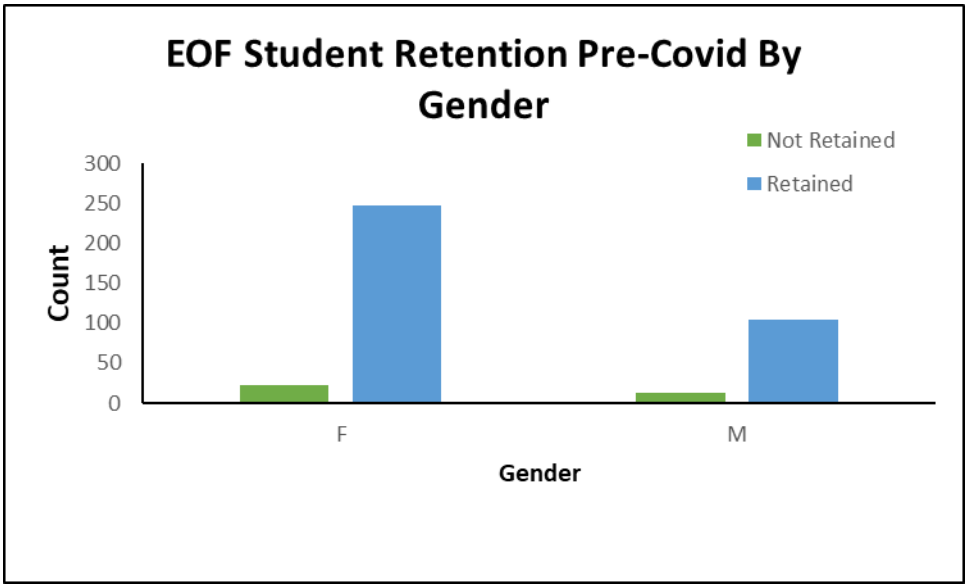


Figure 3.4.2 EOF Student Retention Pre-Covid By Gender

Figure 3.4.3 shows that 16 females did not retain, or 19.3% of the EOF female population post-covid did not retain, and 9 males did not retain, or 29% of the EOF male population post-covid did not retain.

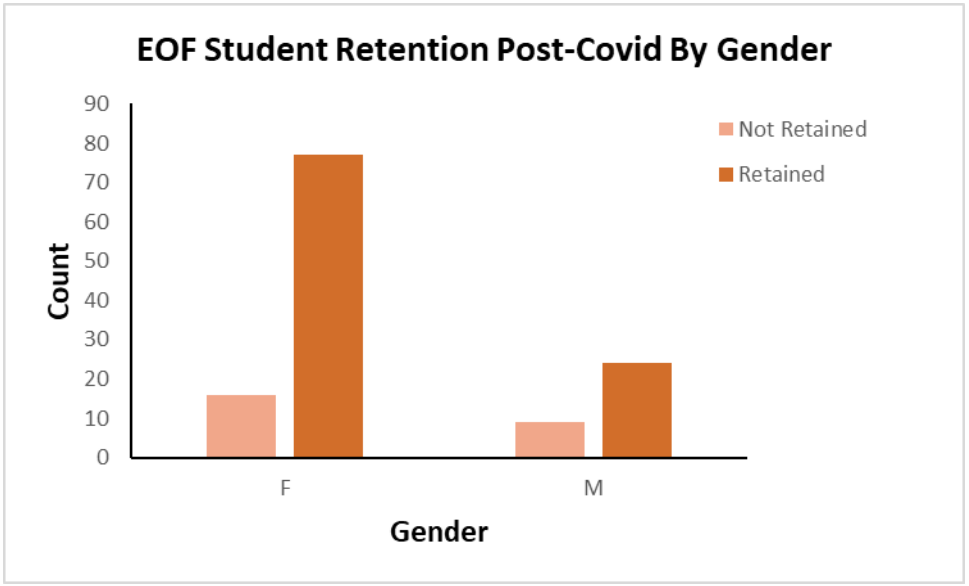


Figure 1.4.3 EOF Student Retention Post-Covid By Gender

The results from this section suggest, based on Figures 3.4.1, 3.4.2, and 3.4.3, that while females have a higher count for attrition, when looking at the percentage of the population per gender, males have a higher percentage rate of attrition than females. Furthermore, while the exact values for the attrition instances for males and females is the lowest post-covid, compared to pre-covid, the attrition percentages are higher as there was a 11.5% increase in attrition rates for the female EOF population, and an 18% increase in attrition rates for the male EOF population.

Section 3.5 Examining Student Retention and Class

The distribution of EOF student retention and attrition is now studied through the lens of class, where students can fall into one of the following four categories- freshman, sophomore, junior or senior. Figure 3.5.1 shows that the top two classes with the highest attrition occurrences were freshman with 56 students, or 11.8% of the EOF freshman class not retaining and sophomores, with 3 students, or 17.6 % of the EOF sophomore class not retaining.

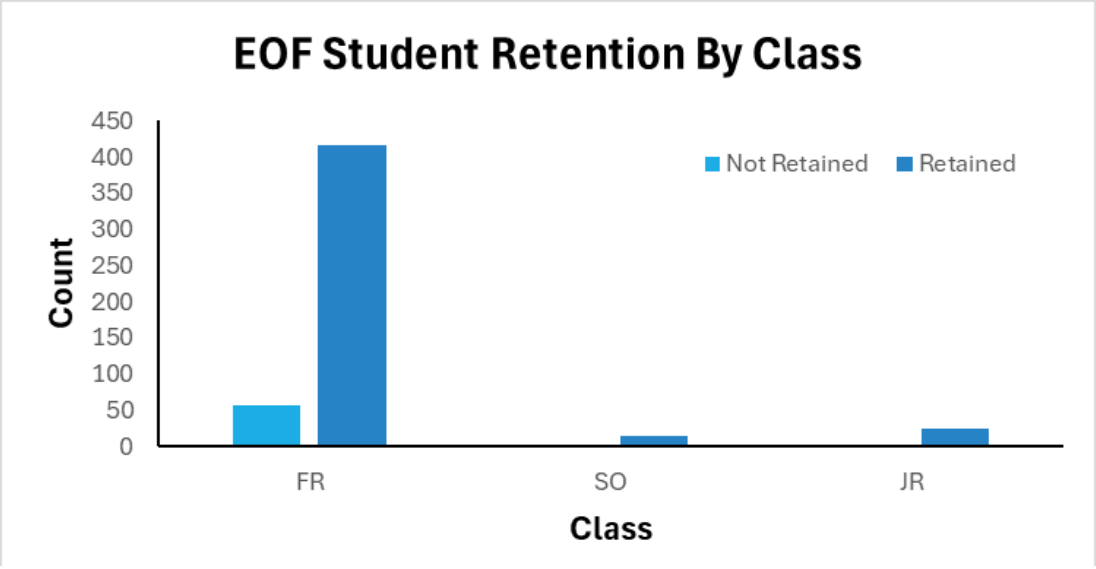


Figure 3.5.1 EOF Student Retention By Class

Pre-covid, Figure 3.5.2 shows that the top two classes with the highest attrition occurrences were freshman with 34 students, or 9.19% of the EOF freshman class pre-covid not retaining and sophomores, with 1 student, or 8.33% of the EOF sophomore class pre-covid not retaining.

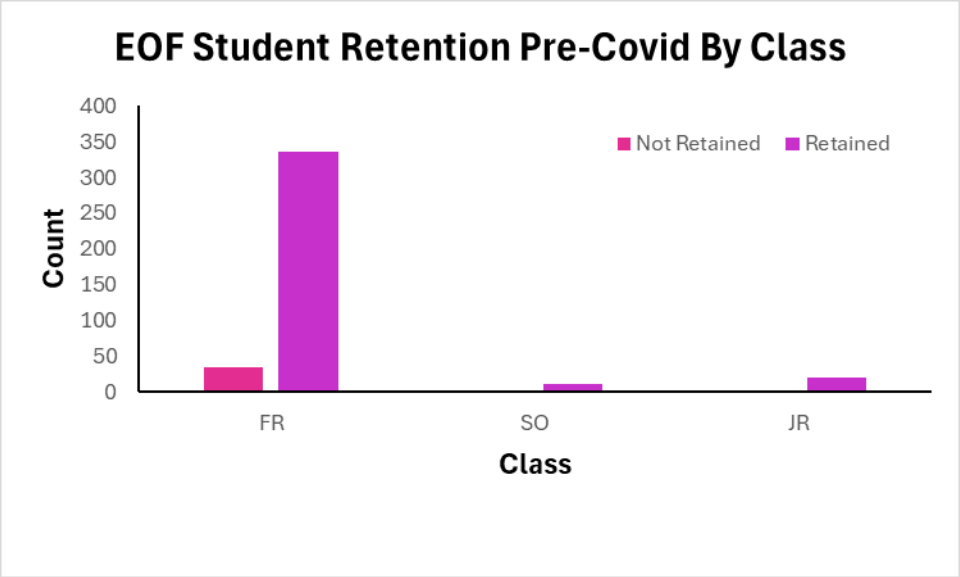


Figure 3.5.2 EOF Student Retention Pre-Covid By Class

Post-covid, Figure 3.5.3 shows that the top two classes with the highest attrition occurrences were freshmen, with 22 students, or 21.4% of the EOF freshman class post-covid not retaining, and sophomores, with 2 students, or 40.0% of the EOF sophomore class post-covid not retaining.

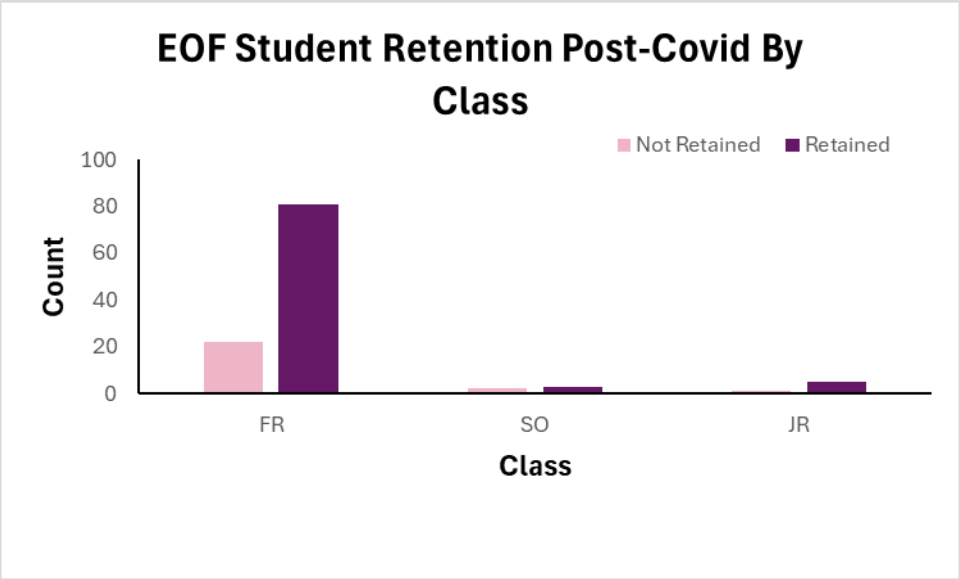


Figure 3.5.3 EOF Student Retention Post-Covid By Class

Overall, the results as shown in Figures 3.5.1, 3.5.2, and 3.5.3, suggest that students who belong to the freshman or sophomore class have higher attrition rates, especially post-covid. When comparing the attrition rates between these two classes, as a population sophomores had higher attrition rates in two out of three of the analyses, specifically for all EOF students and all EOF students post-covid.

Section 3.6 Examining Student Retention and Styp Code

The distribution of EOF student retention and attrition is now explored from the perspective of styp-code, which classifies a student as a first-time new student, a transfer student, a continuing student, or a non-matriculated student. Figure 3.6.1 shows that students who are categorized as new first-time students, where it is usually their first semester at Ramapo, have the highest occurrences of student attrition, with 55 students or 11.7% of the new student population. Subsequently, transfer students have 4 recorded instances of student attrition, or 9.5% of the transfer student population.

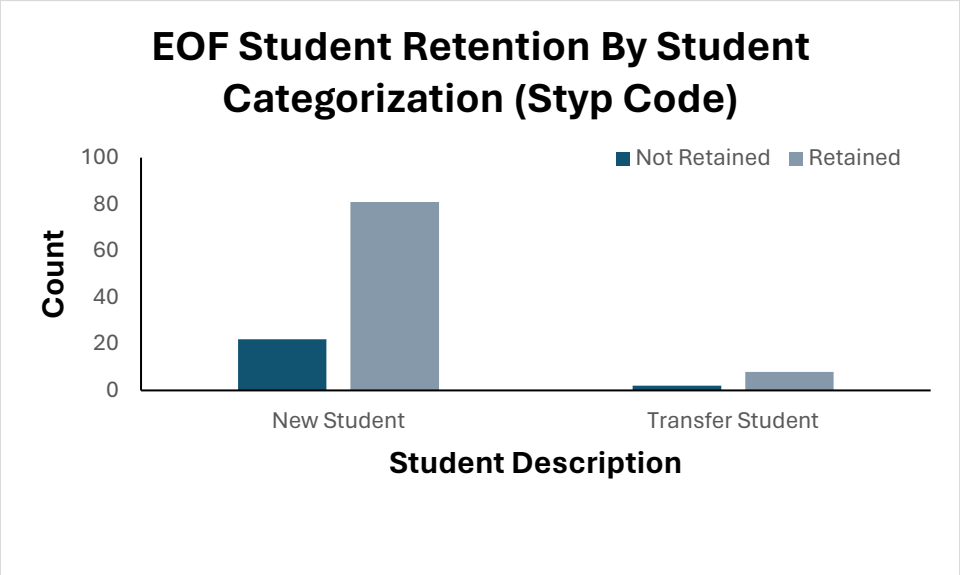


Figure 3.6.1 EOF Student Retention By Styp Code

Similarly to the results shown in Figure 3.5.1, for all EOF students pre-covid, new first-time students had the highest attrition rates, with 33 students, or 8.9% of the new first-time student population not retaining. For transfer students, 2 students or 6.3% of the transfer student population did not retain. This is shown in Figure 3.6.2.

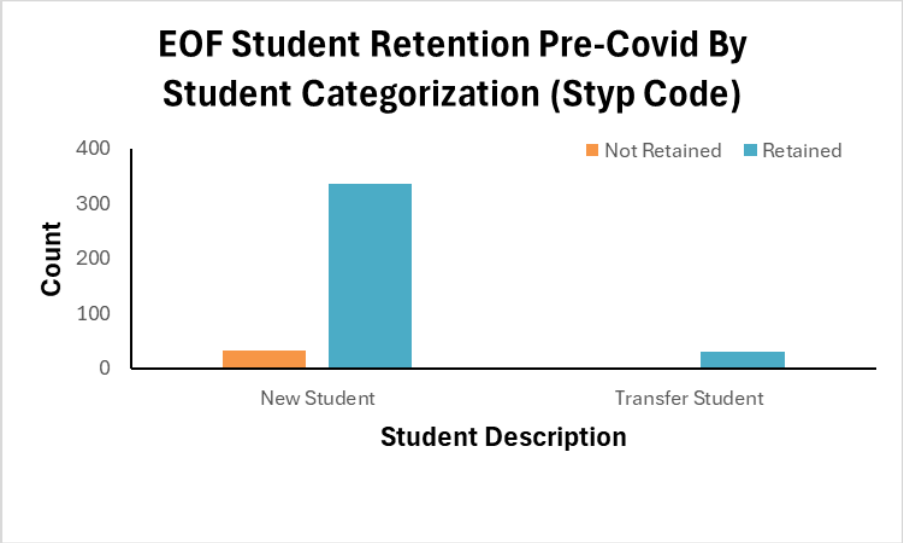


Figure 3.6.2 EOF Student Retention Pre-Covid By Styp Code

The results are consistent for the post-covid analysis which is shown in Figure 3.6.3. Twenty-two new first-time students, or 21.4% of the new first-time student population did not retain. For transfer students, 2 students or 20% of the transfer student population did not retain.

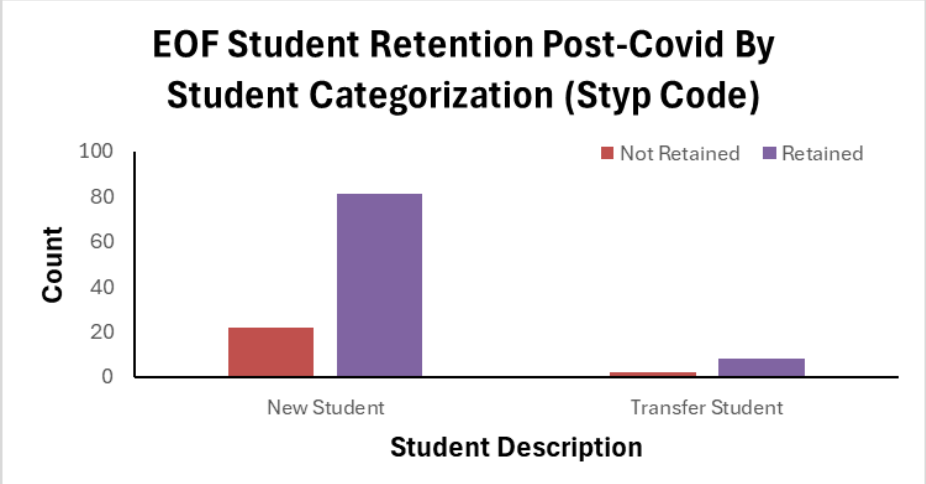


Figure 3.6.3 EOF Student Retention Post-Covid By Styp Code

For all three analyses regarding student retention and student categorization, the results as shown in Figures 3.6.1, 3.6.2, and 3.6.3, suggest that new first-time students tend to have higher attrition rates than transfer students.

Section 3.7 Examining Student Retention and Residency Status

Within this section, the relationship between EOF student retention and residency status, whether they live in one of the residence halls at Ramapo or are a commuter, are examined. The results are consistent for all three subdivisions of this study as EOF students who live on campus, residents, have both a higher retention and attrition rate than EOF commuters, students who do not live on campus. Figure 3.7.1 shows that when considering the entire EOF student population, 50 residents, or 12.3% of residents' population did not retain, and 10 commuters, or 9.1% of the commuter population did not retain.

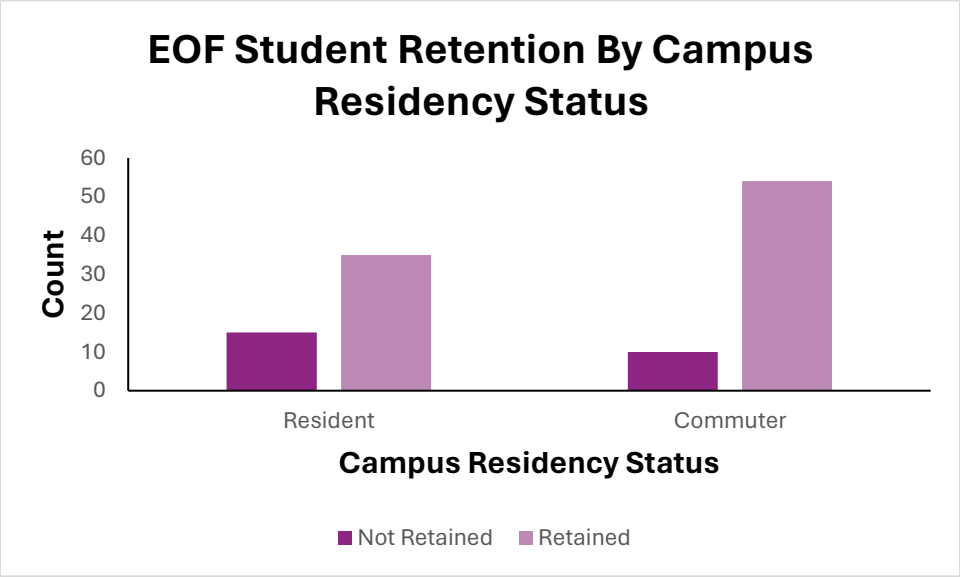


Figure 3.7.1 EOF Student Retention By Campus Residency Status

The results for the EOF students pre-covid were slightly different as 35 students or 9.9% of the residents' population pre-covid did not retain, whereas for students who commute, 0 students did not retain, which is shown in Figure 3.7.2.

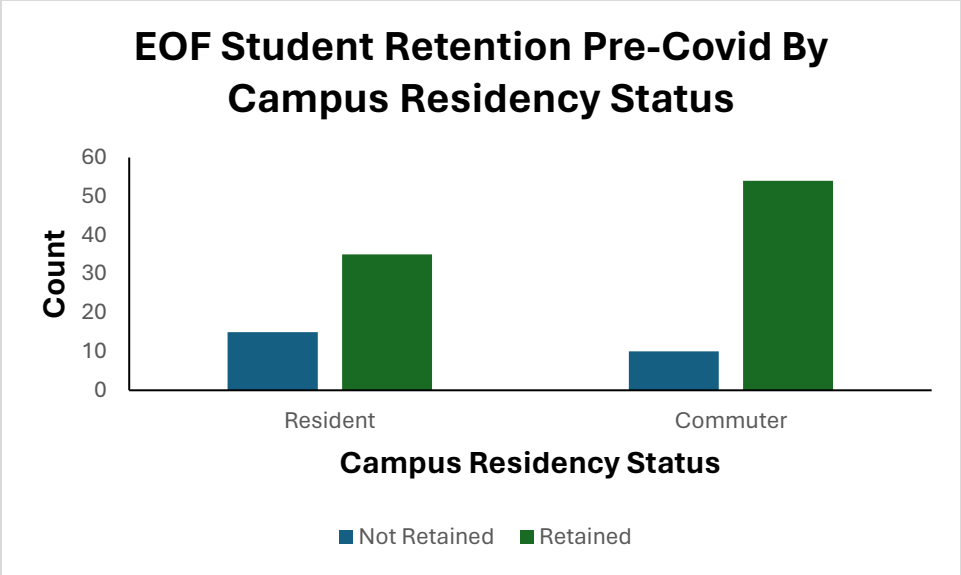


Figure 3.7.2 EOF Student Retention Pre-Covid By Campus Residency Status

For EOF students post-covid, 15 students or 30% of the EOF residents' population post-covid did not retain, but for commuters, 10 students, or 15.6% of the EOF commuter population post-covid did not retain, which is shown in Figure 3.7.3.

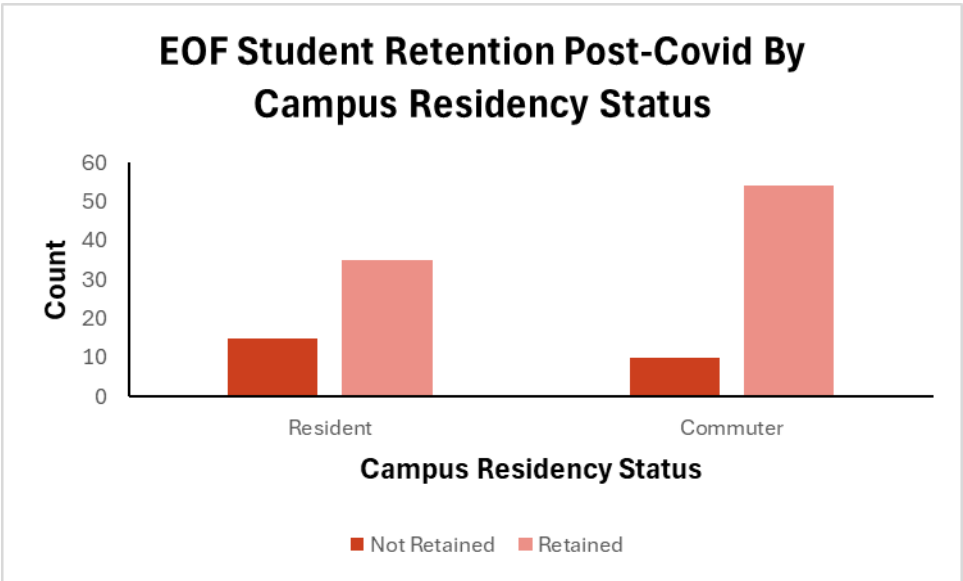


Figure 3.7.3 EOF Student Retention Post-Covid By Campus Residency Status

These results, as shown in Figures 3.7.1, 3.7.2, and 3.7.3, suggest that EOF residents have both higher retention and attrition rates than EOF commuters do. When the analysis is broken down into pre-covid and post-covid, there was a 20.1% increase in resident students who did not retain, and a 15.6% increase in commuter students who did not retain, post-covid.

Section 3.8 Examining Student Retention and Average GPA (Cumulative & Term)

EOF student retention and attrition is now examined from the perspective of term GPA, and cumulative GPA. Table 3.8 provides the average cumulative GPA, and the average term GPA for both students who have retained, denoted R, and students who have not retained, denoted NR, for all three subdivisions of this analysis – all EOF students, all EOF students pre-covid, and all EOF students post-covid.

Table 3.8 Examining Term GPA and Cumulative GPA For All EOF Students Based on Retention

	Avg Term GPA (R)	Avg Cum GPA (R)	Avg Term GPA (NR)	Avg Cum GPA (NR)
All EOF Students	2.76	2.9	1.65	2.15
EOF Students Pre-Covid	2.73	2.88	1.52	2.22
EOF Students Post-Covid	2.92	2.99	1.83	2.04

As shown in Table 3.8, there is an average difference of 1.14 grade points for average term GPA between EOF students who did and did not retain, and there is an average difference of 0.79 grade points for average cumulative GPA, between EOF students who did and did not retain. EOF students who did retain had a higher average term GPA of 0.19 grade points and a higher average cumulative GPA by 0.11 grade points post-covid than they did pre-covid. However, for EOF students who did not retain, their average term GPA was higher post-COVID,

by 0.31 grade points, whereas their average cumulative GPA was lower by 0.18 grade points, post-COVID.

Overall, the average term GPA (R), average cumulative GPA (R), and average term GPA (NR) was the highest post-covid, which can be attributed to the different grading options available during Spring 2020. During this semester at Ramapo College students were given the option to receive a Pass (P) or Fail (F) instead of a letter grade. For the students who would have earned letter grades of C and D, they elected for a pass (P), which circumvented a lower GPA.

Chapter 4 EOF Population Report Card

This chapter of the analysis provides information regarding the average term GPA over time, grade distribution, and the courses EOF students tend to struggle in. For this chapter, unlike previous ones, the analysis focuses on all EOF students, regardless of whether they retained or not, over the time span of Fall 2013 to Spring 2023, and there is no distinction made between pre-covid and post-covid.

Section 4.1 EOF Average Term GPA

Figure 4.1.1 shows the average term GPA for all EOF students, regardless of whether they retained or not from Fall 2013 through Spring 2023.

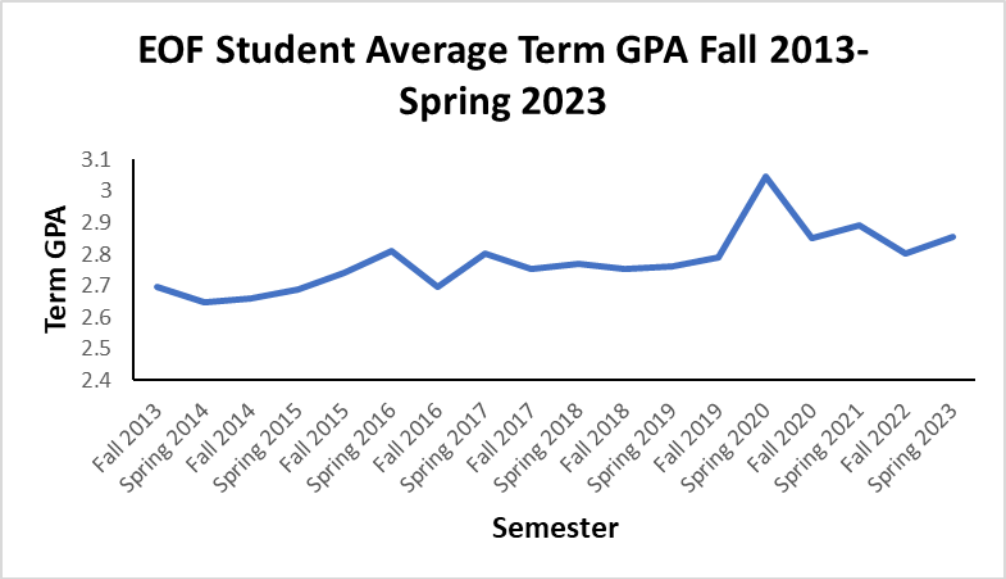


Figure 4.1.1 EOF Student Average Term GPA Fall 2013-Spring 2023

Figure 4.1.1 shows that the average term GPA was the lowest at 2.65, in Spring 2014, and had another significant drop to 2.7, in Fall 2016. Conversely, the average term GPA was the highest in Spring 2020, which is when the covid-19 pandemic surged. This is unsurprising as during this semester, students were offered the opportunity to receive a Pass or Fail, instead of a

letter grade. Thus, the students who would potentially earn a C or D, may have elected to get a P, which would increase the overall GPA.

In order to provide a complete picture of the grade distribution for EOF students, while still providing the EOF department with the grade distribution they are most interested, that is, students receiving a D+, D, F, P, or W, the analysis was broken down into Figures 4.1.2, and 4.1.3.

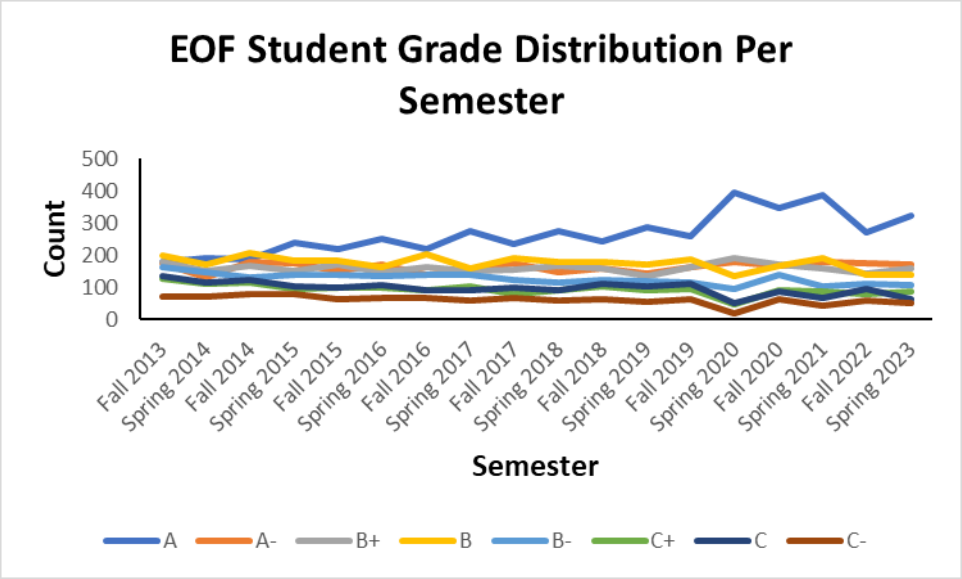


Figure 4.1.2 EOF Grade Distribution (A through C-) Per Semester

As seen from Figure 4.1.2, there was an increase in the number of students receiving letter grades of A, A-, and B+, whereas there was a decrease in the number of students receiving the letter grades of B, B-, C+, C, and C-, during Spring 2020. Figure 4.1.2 shows how the covid-19 pandemic affected and impacted the general pre-existing trends of the data. This is also shown in the figure, as there was a significant spike in the number of students who received a P, passing, during the Spring 2020 semester. This is a result of the available option at the time for students to have a pass, P, or fail, F, on their official transcript instead of the traditional letter grade. Furthermore, Figure 4.1.3 shows that there was a consequent increase in the number of

students withdrawing, and failing, whereas there was a decrease in the number of students earning the letter grade of D+, and D.

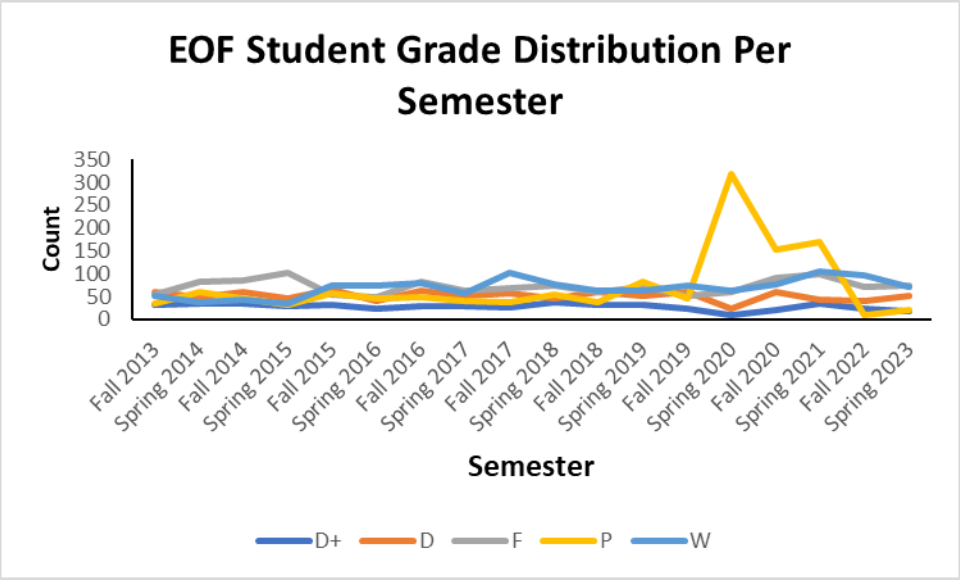


Figure 4.1.3 EOF Grade Distribution (D+ through W) Per Semester

Section 4.2 Determining the Courses Where EOF Students Struggle

Within this section, the focus is on determining the classes and, more broadly, the subject areas that EOF students are struggling in. For those reasons and to increase interpretability, this analysis will specifically be targeted at the top ten classes, and subject areas, which have the highest counts of EOF students receiving the letter grades of D, F, and W. Figure 4.2.1 shows the top 10 courses that had the highest counts of EOF students receiving the letter grade, D.

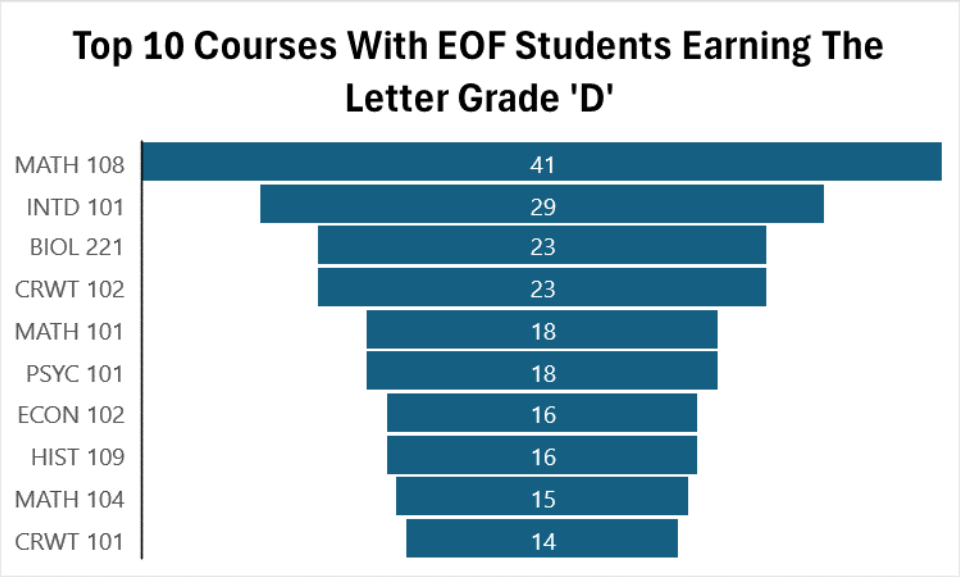


Figure 4.2.1 Top 10 Courses With EOF Students Earing The Letter Grade 'D'

As shown in Figure 4.2.1, the top four courses with the highest instances of students receiving the letter grade D are math 108, interdisciplinary studies 101, biology 221, and critical reading and writing 101, with counts of 41, 29, 23, and 23, respectively. Figure 4.2.2 shows the top 10 courses with students earning the letter grade, F.

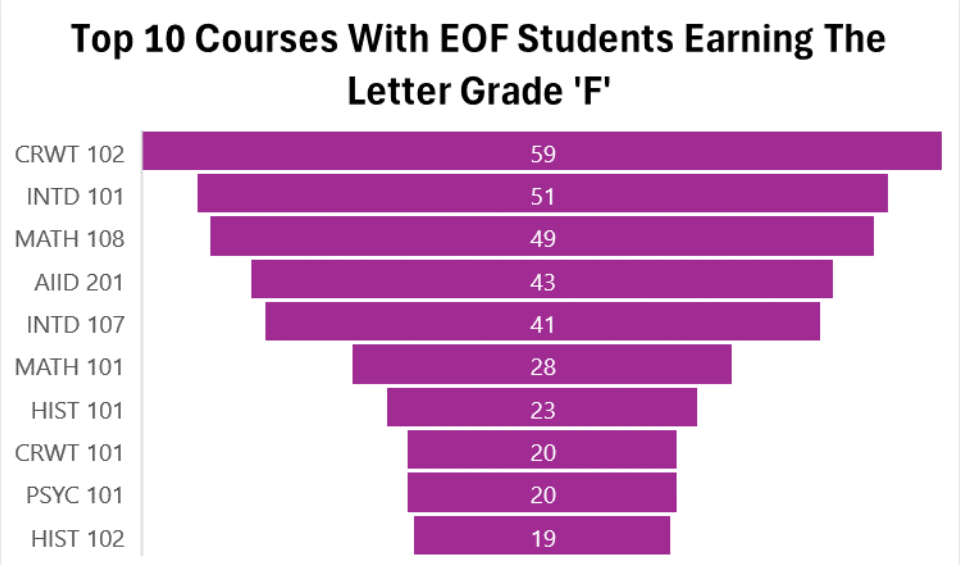


Figure 4.2.2 Top 10 Courses With EOF Students Earning The Letter Grade 'F'

As seen from Figure 4.2.2, the top four courses with the highest number of students receiving the letter grade F are critical reading and writing 102, interdisciplinary studies 101, math 108, and amer/intl interdisciplinary 201, with counts of 59, 51, 49, and 43, respectively. Comparing the top four courses with the highest counts of students receiving the letter grades D and F respectively, more students received F grades than D grades.

Figure 4.2.3 shows the top 10 courses with students earning the letter grade, W, for withdrawal. Three out of the four top courses with the highest instances of students withdrawing are math courses, specifically math 108, math 101, and math 110, with counts of 63, 31, and 28, respectively. Similar to the previous figure, amer/intl interdisciplinary 201, is also within the top four, as it is the second course with the highest number of EOF student withdrawals, 35 precisely.

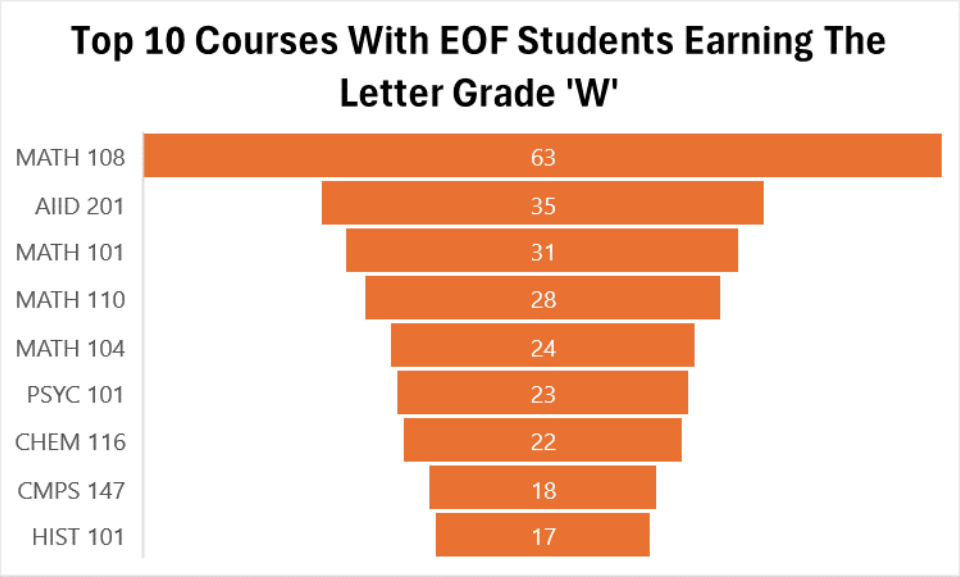


Figure 4.2.3 Top 10 Courses With EOF Students Withdrawing

The results from Figures 4.2.1, 4.2.2, and 4.2.3, demonstrate the thirty classes that EOF student struggle in the most, by specifically looking at the letter grades of D, F, and W. When focusing on the top four classes within each figure, the EOF department should provide

additional resources and support for students enrolled in the following courses- math 108, interdisciplinary study 101, biology 221, critical reading and writing 102, amer/intl interdisciplinary 201, math 101, and math 110.

Section 4.3 Determining Subject Areas Where EOF Students Struggle

Where Section 4.2 detailed the specific courses that EOF students typically struggle in, within this section, the analysis is expanded to look at the subject areas EOF students are struggling in, which is defined as students earning a D, F, or W letter grade. Figure 4.3.1 shows out of the top 10 subject codes that had the highest occurrences of EOF students earning the letter grade D, the top four were math, biology, psychology, and chemistry, with counts of 107, 83, 74, and 57.

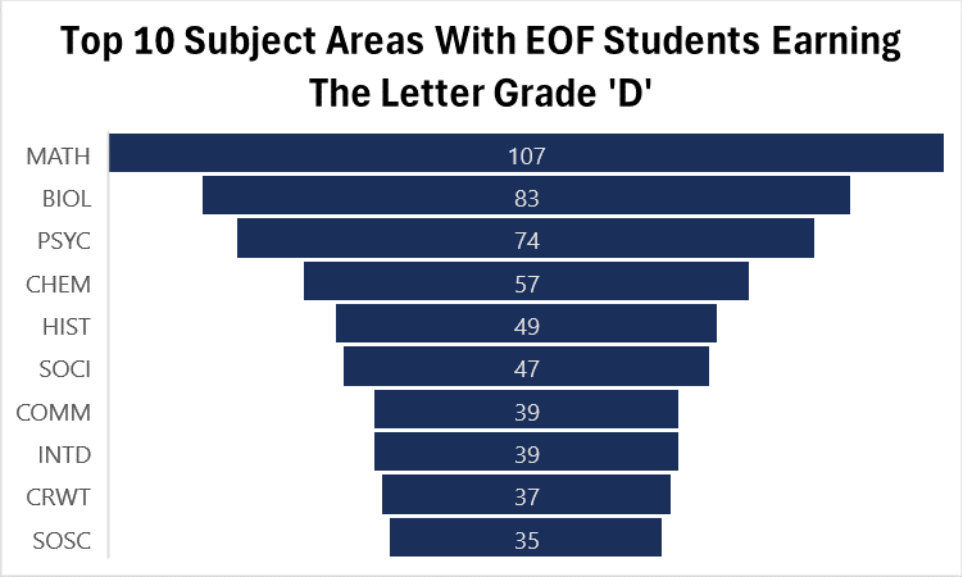


Figure 4.3.1 Top 10 Subject Areas With EOF Students Earning The Letter Grade 'D'

The order of the subject areas for the top subject areas with EOF students receiving the letter grade F is almost identical Figure 4.3.1, for EOF students earning the letter grade F. Similar to the previous result math takes the lead, however instead of biology, interdisciplinary

studies follow, then psychology, and chemistry, with counts of 107, 83, 74, and 57, respectively. These results are depicted in Figure 4.3.2.

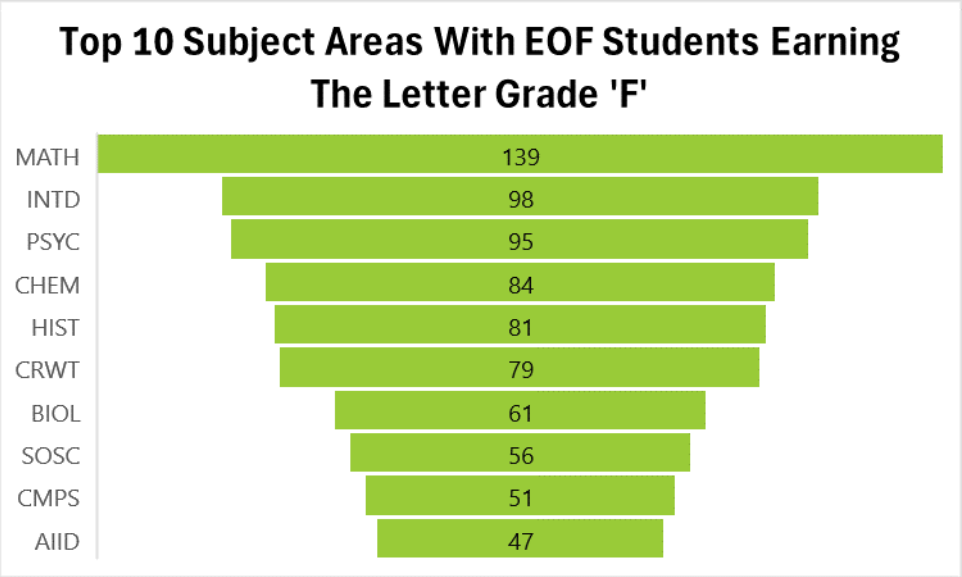


Figure 4.3.2 Top 10 Subject Areas With EOF Students Earning The Letter Grade 'F'

When examining the subject areas with the highest counts of EOF students withdrawing from the class, math is in the lead, followed by psychology, biology, and chemistry, with counts of 231, 111, 94, and 86, respectively. This is shown in Figure 4.3.3.

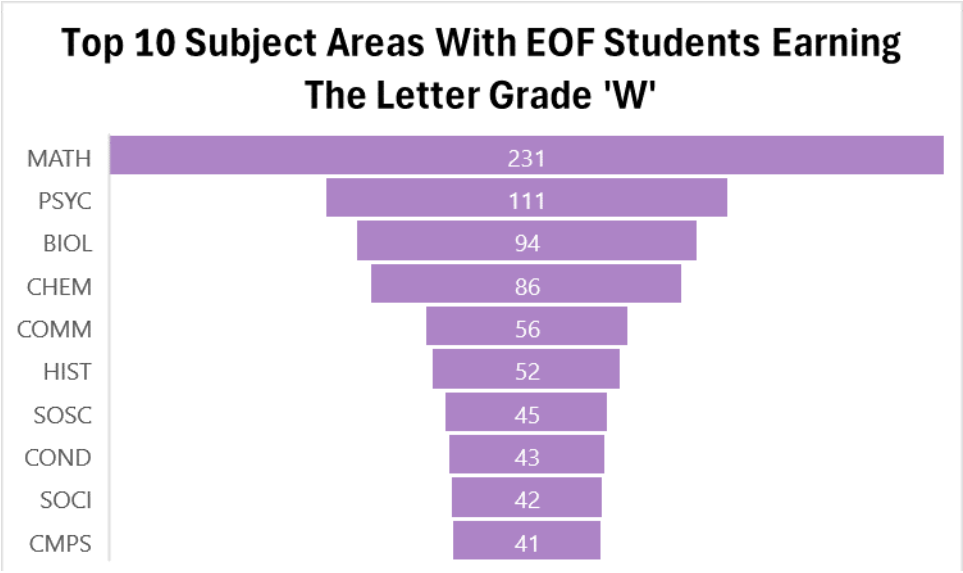


Figure 4.3.3 Top 10 Subject Areas With EOF Students Earning The Letter Grade 'W'

Based on the results from Figures 4.3.1, 4.3.2, and 4.3.3, they suggest that the EOF department may want to provide additional support to students in the following subject areas- math, biology, interdisciplinary studies, psychology, and chemistry.

Chapter 5 A Spotlight on EOF Students In STEM

This chapter focuses on the EOF students who are STEM majors by examining their distribution of retention and attrition rates in general, pre-covid, and post-covid. Similar to the previous chapters, the specific courses and subject areas, which are distinguished through the Ramapo College School of Theoretical and Applied Science classification, that EOF STEM majors struggle in are explored.

Section 5.1 Examining EOF Stem Major Retention and Major

Looking at all of the data from Fall 2013 to Spring 2023, only 83 students, or 16.2% of the EOF student population were STEM majors. Figure 5.1.1 shows the retention of STEM majors based on the available majors, not all of the majors within the School of Theoretical and Applied Science. The biology and computer science majors have the highest occurrences of students not retaining, with counts of 5 and 3, whereas the chemistry, environmental studies, environmental science, and math majors have the lowest retention rates, with counts of 4, 2, 2, and 1, respectively.

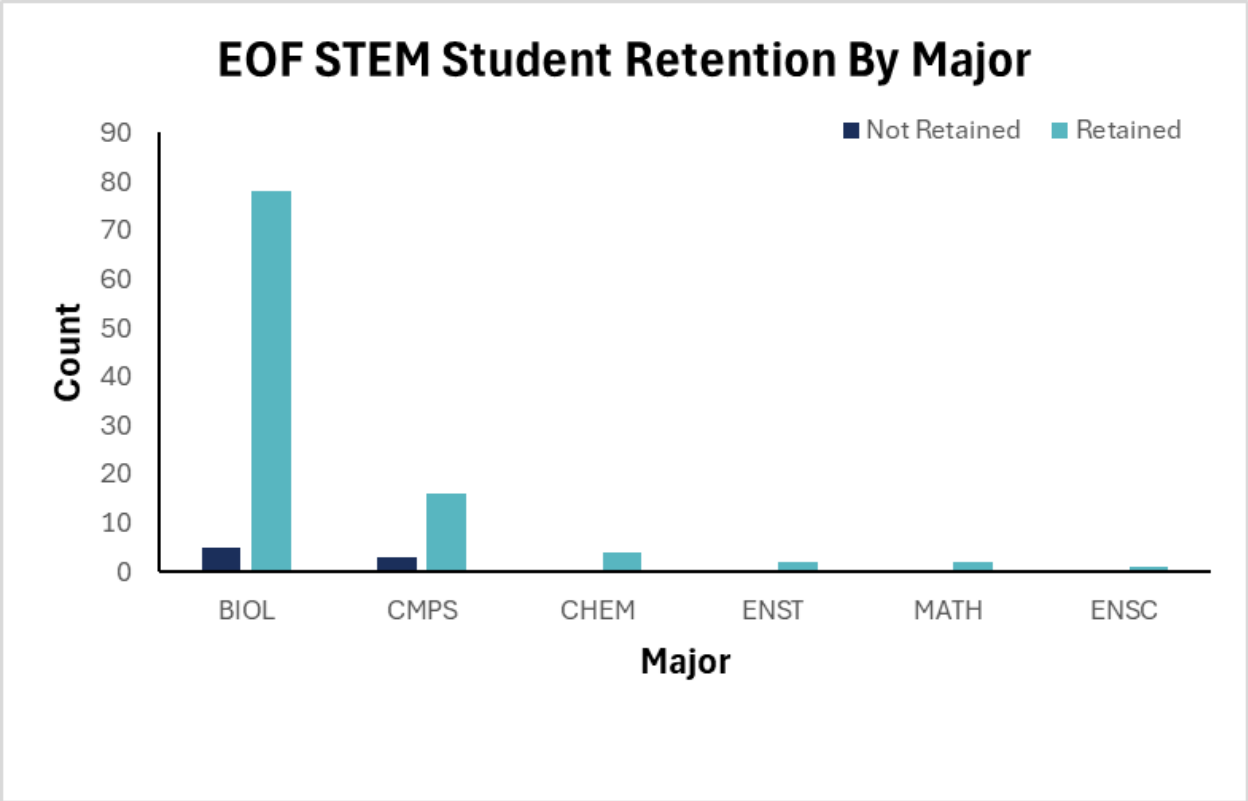


Figure 5.1.1 EOF Stem Student Retention By Major

Breaking the analysis down further, Figure 5.1.2 shows that pre-covid, the majors of biology and computer science, similar to the general analysis, had the highest attrition occurrences with counts of 3, and 2, respectively. The chemistry, environmental studies and math majors had the lowest attrition rates, with students counts of 2, 2, and 1, respectively.

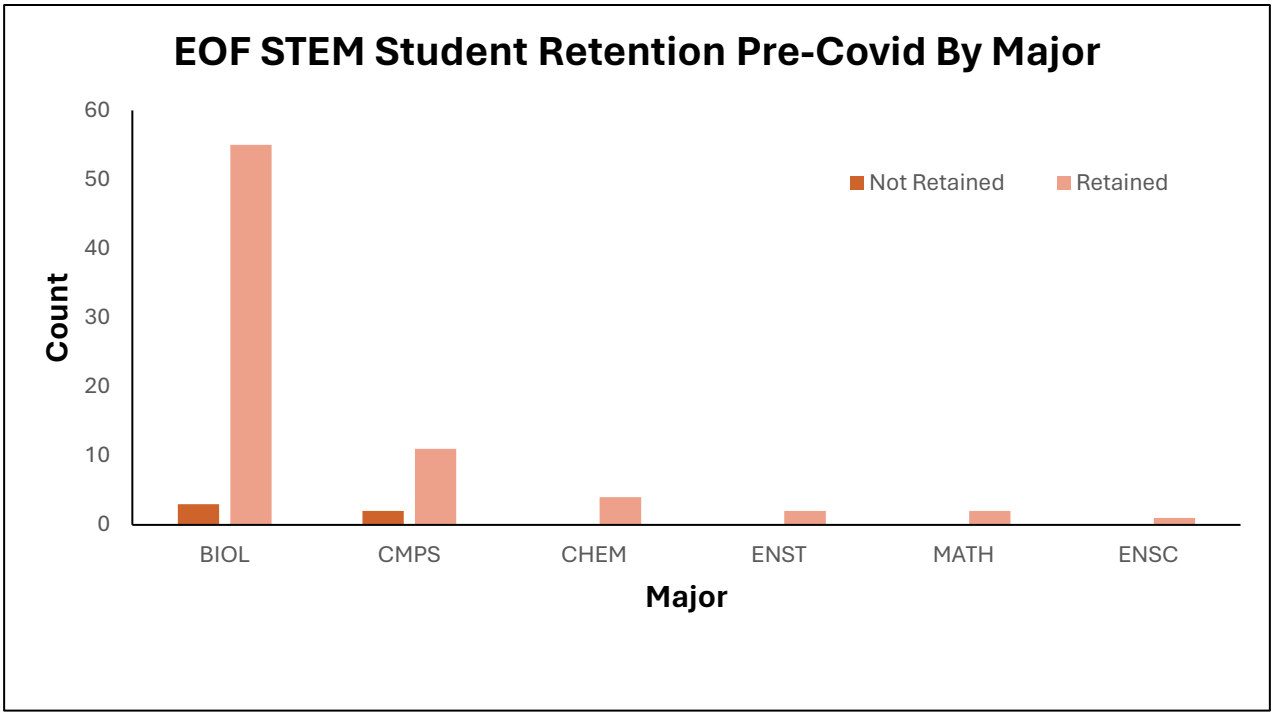


Figure 5.1.2 EOF Stem Student Retention Pre-Covid By Major

Post-covid, the data is only limited to biology and computer science majors which is shown in Figure 5.1.3. Two biology majors, or 8.7% of EOF biology majors did not retain, whereas 1 student or 1.7% of EOF computer science majors did not retain.

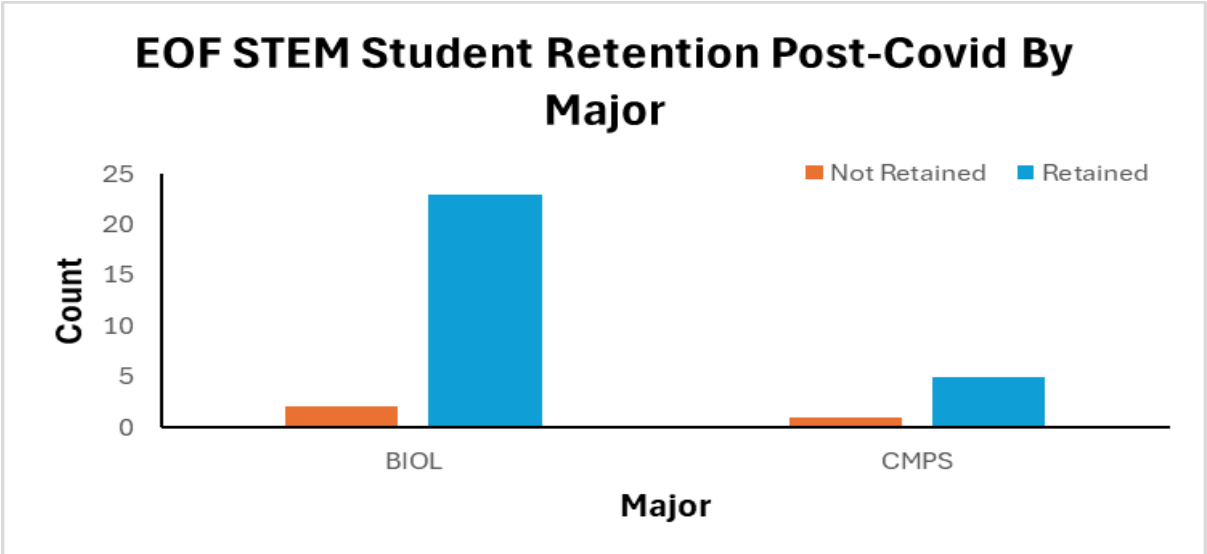


Figure 5.1.3 EOF Stem Student Retention Post-Covid By Major

Based on the results from Figures 5.1.1, 5.1.2, and 5.1.3, the EOF department may want to provide extra support and resources for their students who are majoring in biology or computer science.

Section 5.2 The Courses That EOF STEM Majors Struggle In

For EOF STEM majors, the courses that students struggled with, which are defined as earning a letter grade D, F, or W, for the semester, are determined within this section. Figure 5.2.1 shows the distribution of students earning the letter grade D, and three out of the top four courses are all math courses, that is, math 108, math 101, and math 104, with 41,18, and 15 students earning a D for the semester, respectively. The second course with the highest frequency of students earning the letter grade D is biology 221, with a count of 23 students. Interestingly, 50% of the courses that students are earning a D in are math courses, which is also depicted in Figure 5.2.1.

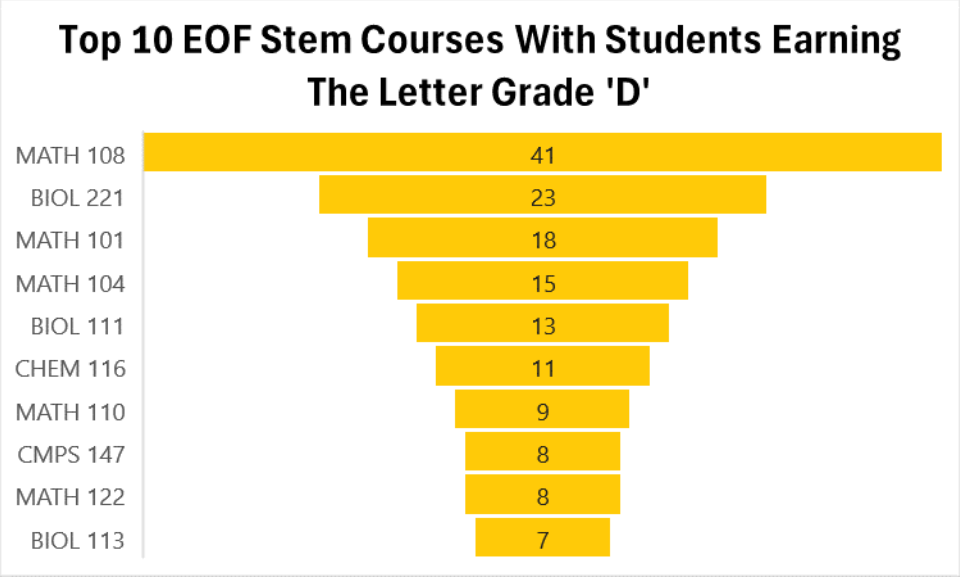


Figure 5.2.1 Top 10 EOF Stem Courses With Students Earning The Letter Grade 'D'

Similar to the distribution of EOF STEM majors earning the letter grade D, for students earning the letter grade F in the semester, three out of the top four courses are all math courses, that is, math 108, math 101, and math 110, with 49, 28, and 17 students earning an F for the semester, respectively. The third course with the highest frequency of students earning the letter grade F is chemistry 116, with a count of 17 students. Also similar to Figure 5.2.1, within Figure 5.2.2, 50% of the courses where students are earning the letter grade F are math courses.

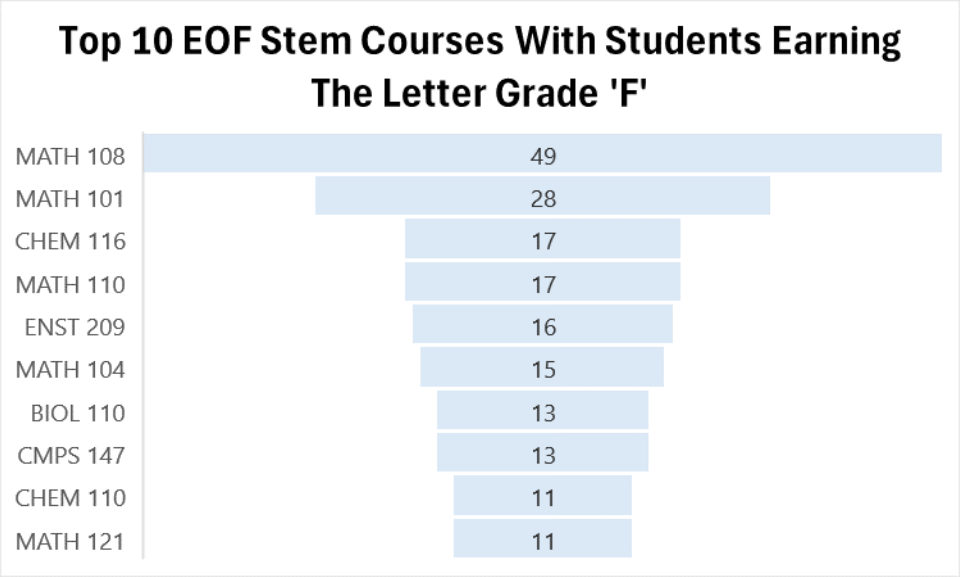


Figure 5.2.2 Top 10 EOF Stem Courses With Students Earning The Letter Grade 'F'

Unlike the previous two letter grade distributions of EOF stem majors earning the letter grades D, and F, as shown in Figures 5.2.1, and 5.2.2, respectively, for students withdrawing, the top four courses are all math courses, specifically, math 108, math 101, math 110, and math 104, with 63, 31, 28, and 24 students withdrawing. While 50% of the courses that students earned D's and F's for were math classes, only 40% of the courses that students withdrew from were math classes. This is demonstrated in Figure 5.2.3.

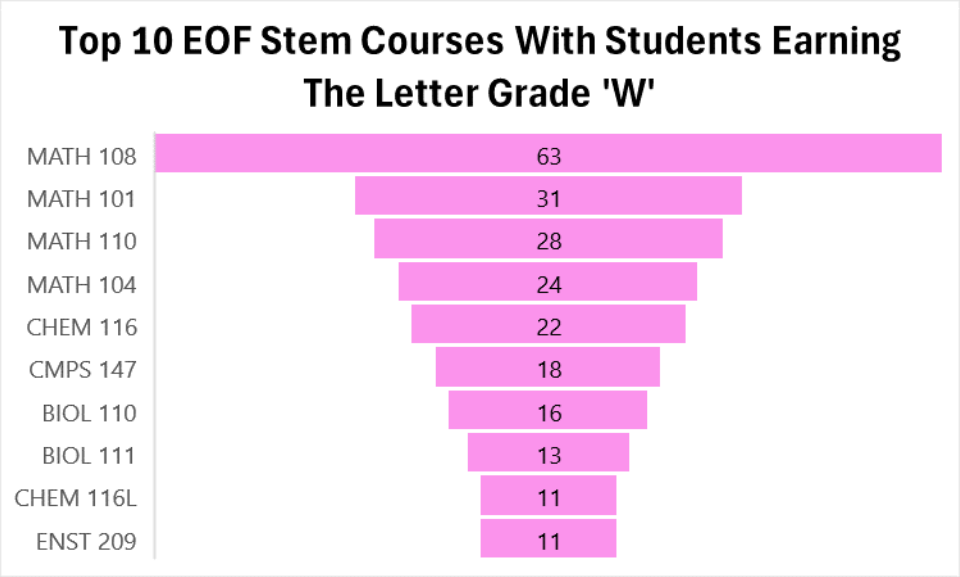


Figure 5.2.3 Top 10 EOF Stem Courses With Students Withdrawing

Based on the top four results from Figures 5.2.1, 5.2.2, and 5.2.3, the EOF department should provide additional resources and support to students who are enrolled in the following STEM courses – math 108, biology 221, math 101, math 104, chemistry 116, and math 110.

Section 5.3 The Subject Areas That EOF STEM Majors Struggle In

While Section 5.2 focused on the specific stem courses that EOF students had a challenging time in, the analysis is now generalized to STEM subject areas. Figure 5.3.1 shows the top 10 stem subject areas where students earned a letter grade D for the semester. Based on the results from the previous section, it is unsurprising that Math has the highest frequency, followed by biology, chemistry, and computer science, with 107, 83, 57, and 26 students earning a D for the semester, respectively.

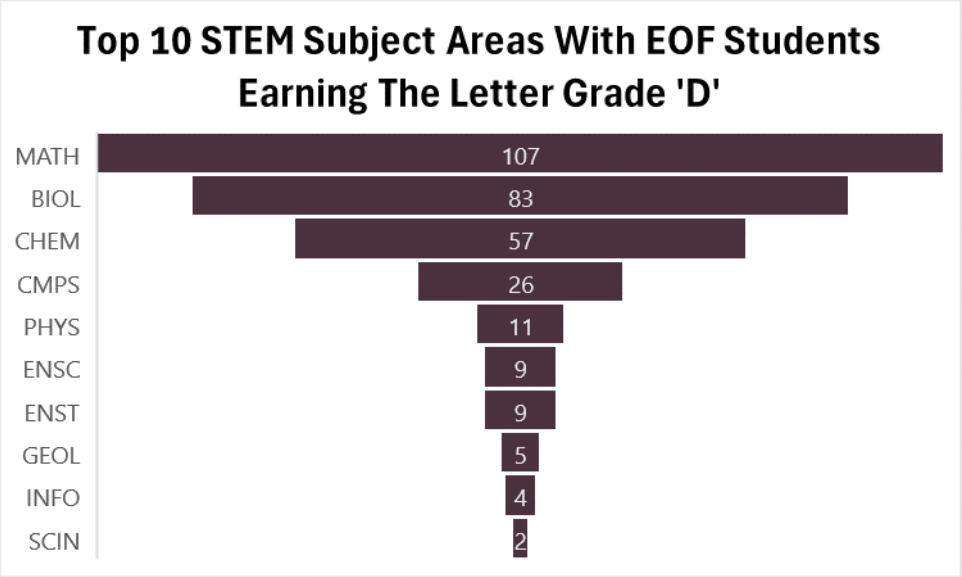


Figure 5.3.1 Top 10 Stem Subject Areas With EOF Students Earning The Letter grade 'D.'

The distribution of students earning the letter grade F in STEM subject areas is very similar to the distribution of students earning the letter grade D, as discussed, however chemistry and biology have switched positions. Math is in first place again, followed by chemistry, biology, and computer science, with 139, 84, 61, and 51 students earning an F for the semester in their STEM subject area. This is depicted in Figure 5.3.2.

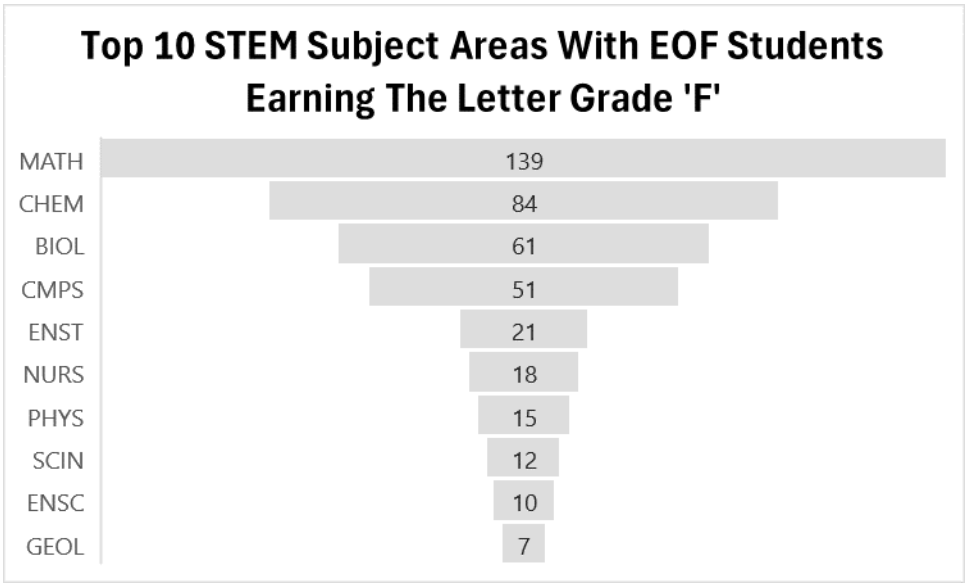


Figure 5.3.2 Top 10 Stem Subject Areas With EOF Students Earning The Letter Grade 'F'

The distribution of students withdrawing from a STEM subject area during the semester is the same as the number of students earning an F for the semester, however the values are different. Looking at the top four, math, chemistry, biology, and computer science have the highest frequency of students withdrawing, with 139, 84, 61, and 51 instances, respectively. This is shown in Figure 5.3.3.

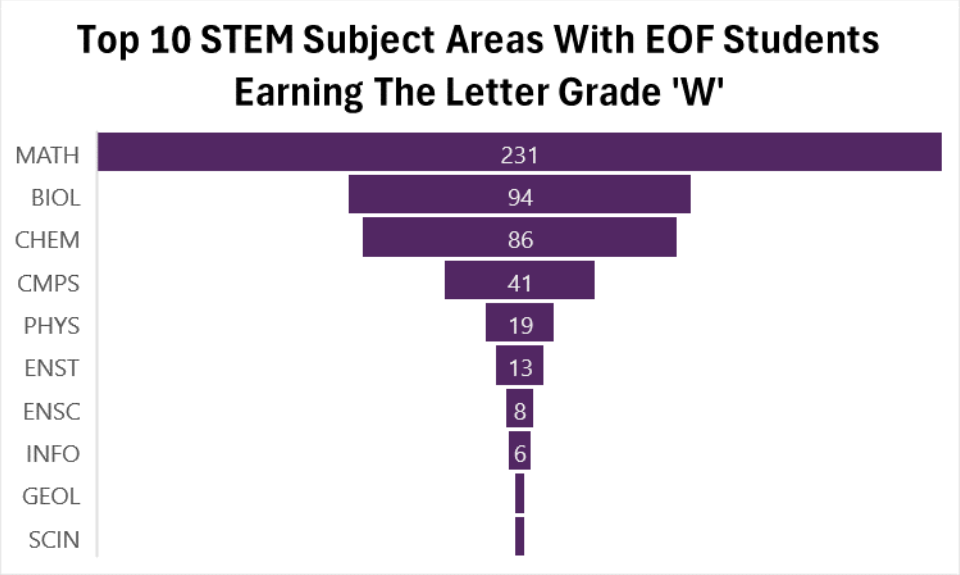


Figure 5.3.3 Top 10 Stem Subject Areas With EOF Students Withdrawing

Overall, based on the results from Figures 5.3.1, 5.3.2, and 5.3.3, it suggests that the EOF department may want to provide additional support to students who are enrolled in math, biology, chemistry, and computer science courses.

Chapter 6 Clustering the EOF Population

Similar to the approach shown in previous chapters, the EOF population is clustered in three segments- all EOF students, all EOF students pre-covid, and all EOF students post-covid, using the k-means method. Within each section, principal component analysis was applied, and by plotting the principal components against their associated percentage of variance explained, it

was determined that the first 2 principal components contain most of the variation within the data.

Section 6.1 Clustering All EOF Students

By iterating through the associated silhouette scores for the number of clusters within the range 2 through 10, the silhouette score for 2 clusters was the highest with a value of 0.55, so the EOF population was sorted into group 0, or group 1. This clustering, along with each group's respective centroid is shown Figure 6.1.1.

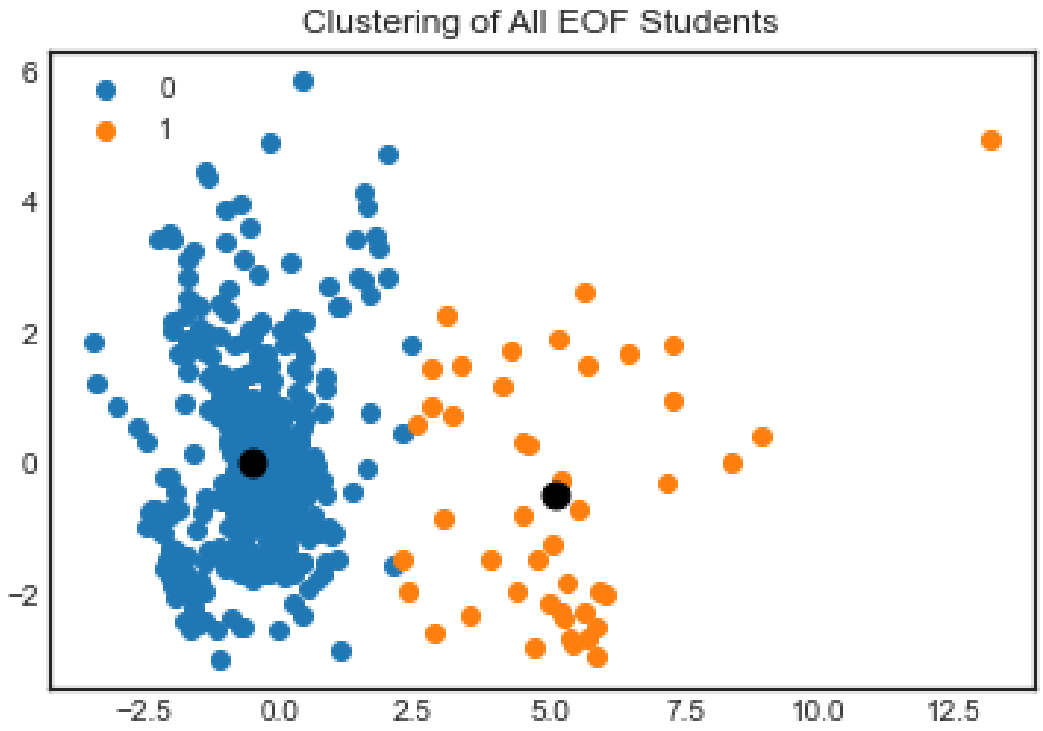


Figure 6.1.1 All EOF Students Clustering

While Figure 6.1.1 provides a nice visual representation of the clusters and their centroids, it does not provide insight about how many or the types of students within the group. Figure 6.1.2 shows the disproportionate distribution of students within each group, where cluster 0 contains the majority of students, 470, followed by cluster 1 with 45 students.

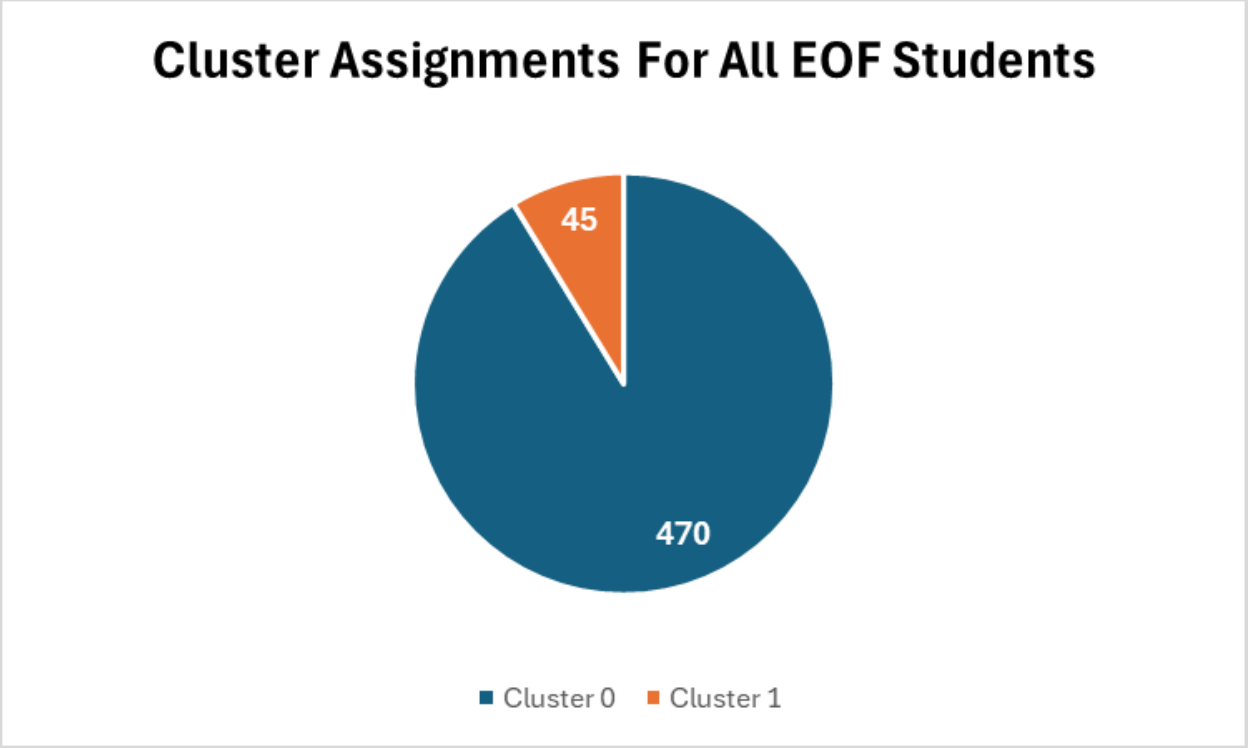


Figure 6.1.2 All EOF Students Clustering

This analysis is expanded to determine the distribution of students based on school as designated by Ramapo. Cluster 0 contains the largest number of students from all five schools, Anisfield School of Business (SB), School of Social Science and Human Services (SS), School of Contemporary Arts (CA), School of Theoretical and Applied Science (TS), and School of Humanities and Global Studies (HG), with 78, 93, 57, 174, and 18, whereas cluster 1 has 10, 18, 5, 9, and 1 students within each school, respectively. These results are shown in Figure 6.1.3.

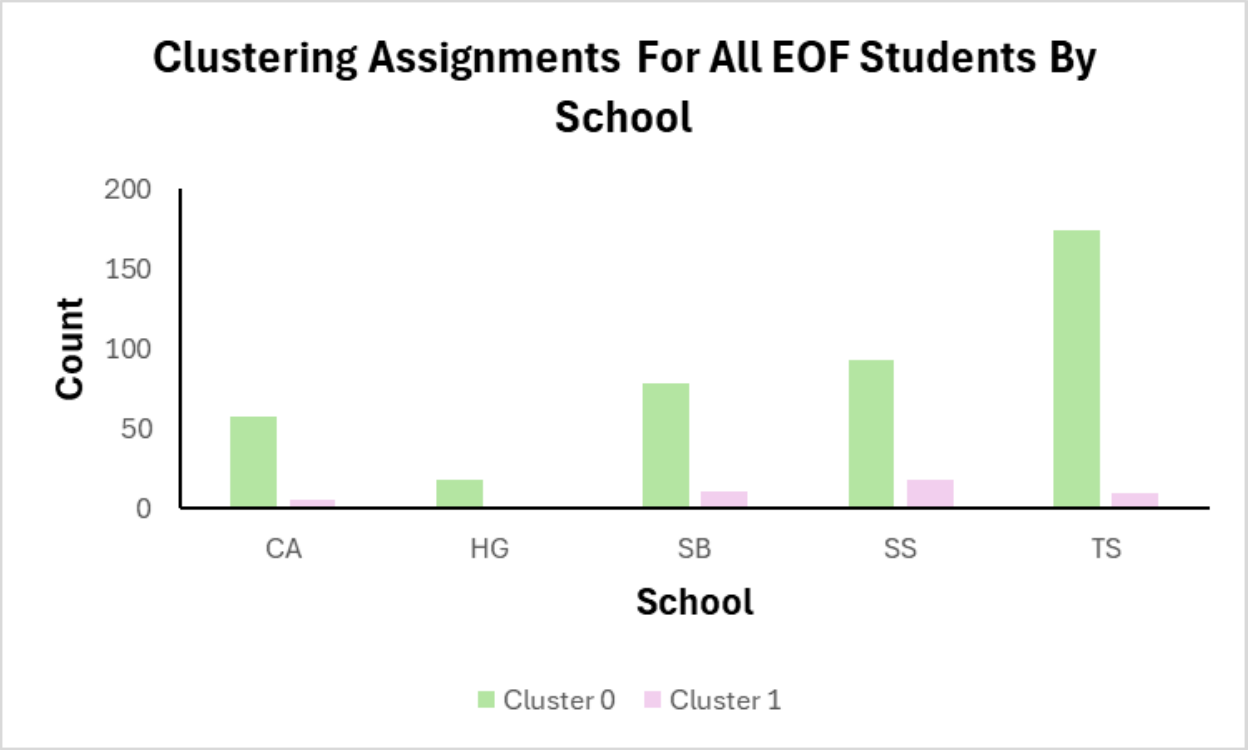


Figure 6.1.3 Clustering Assignments For All EOF Students By School

The distribution of students based on residency status, commuter or on-campus resident, for each cluster is now analyzed. Figure 6.1.4 shows that cluster 0 has the highest number of commuters, 83 students, and residents, 387 students. Whereas cluster 1 has 27 commuters and 18 residents.

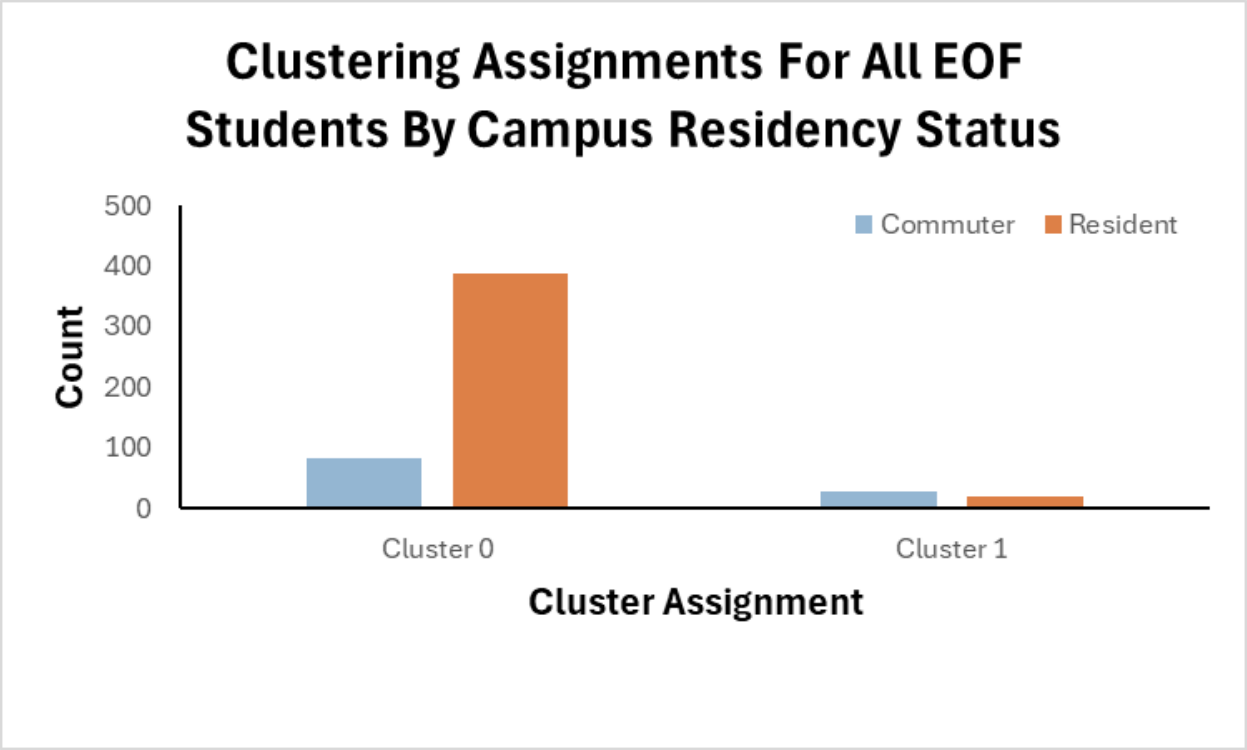


Figure 6.1.4 Clustering Assignments For All EOF Students By Residency Status

When examining the academic performance of students within the cluster, Table 6.1.1 shows that cluster 1 has the highest average cumulative GPA of 2.95, and cluster 0 had the lowest average GPA of 2.8.

Table 6.1.1 Average Cumulative GPA Per Cluster For All EOF Students

Cluster Assignment	Average Cumulative GPA
0	2.797943
1	2.947667

Finally, Figure 6.1.5 shows retention rates amongst both clusters. Cluster 0 has the highest number of students not retaining, 55 or 11.7% of the cluster 0 population. However, even though cluster 1 only has 5 students not retaining, the population of 11.1% of cluster 1 students not retaining is very similar to the cluster 0 attrition rate.

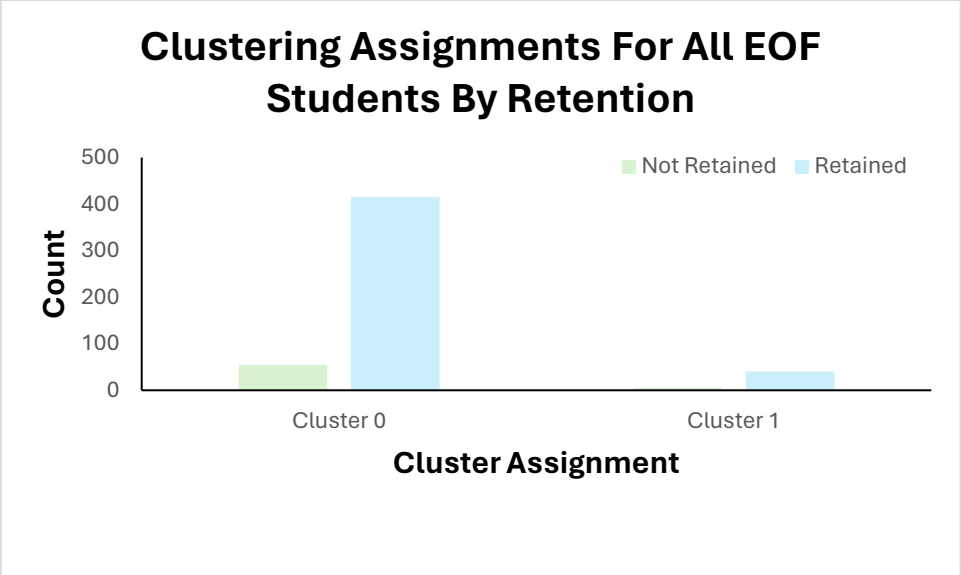


Figure 6.1.5 Clustering Assignments For All EOF Students By Retention

Section 6.2 Clustering All EOF Students Pre-Covid

By iterating through the associated silhouette scores for the number of clusters within the range 2 through 10, the silhouette score for 2 clusters was the highest with a value of 0.56, so the EOF population pre-covid was sorted into group 0 or group 1. This clustering, along with each group's respective centroid is shown in Figure 6.2.1.

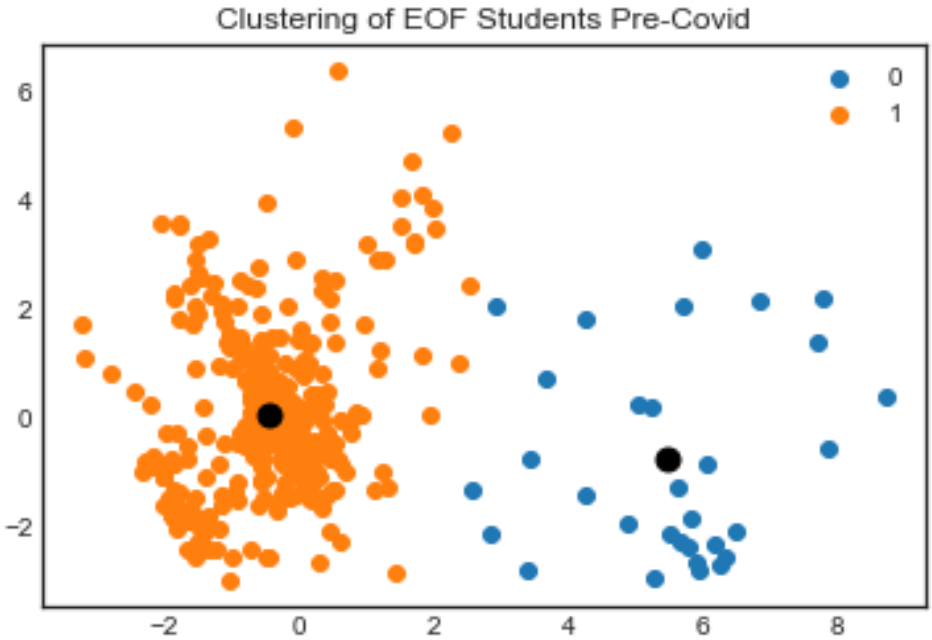


Figure 6.2.1 All EOF Students Clustering Pre-Covid

Figure 6.1.2 shows the disproportionate distribution of students within each group, where cluster 0 contains the majority of students, 370, followed by cluster 1 with 31 students.

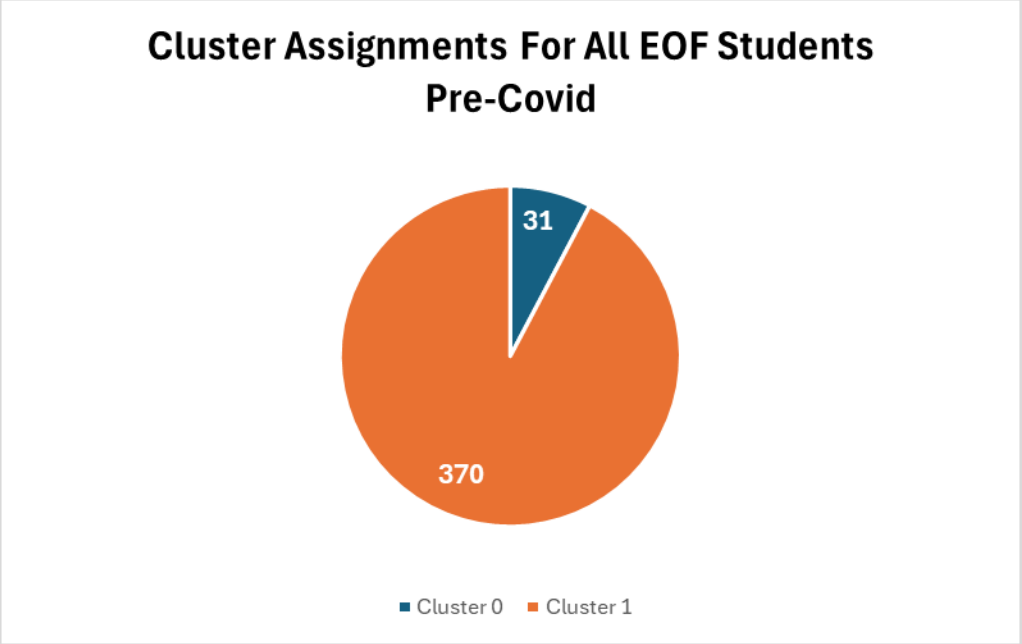


Figure 6.2.2 Cluster Assignments For All EOF Students Pre-Covid

Exploring the distribution of EOF students pre-covid by school, cluster 1 contains the largest number of students from all five schools, Anisfield School of Business (SB), School of Social Science and Human Services (SS), School of Contemporary Arts (CA), School of Theoretical and Applied Science (TS), and School of Humanities and Global Studies (HG), with 59, 75, 47, 127, and 16 students, whereas cluster 1 has 7, 14, 3, 5, and 0 students within each school, respectively. These results are shown in Figure 6.2.3.

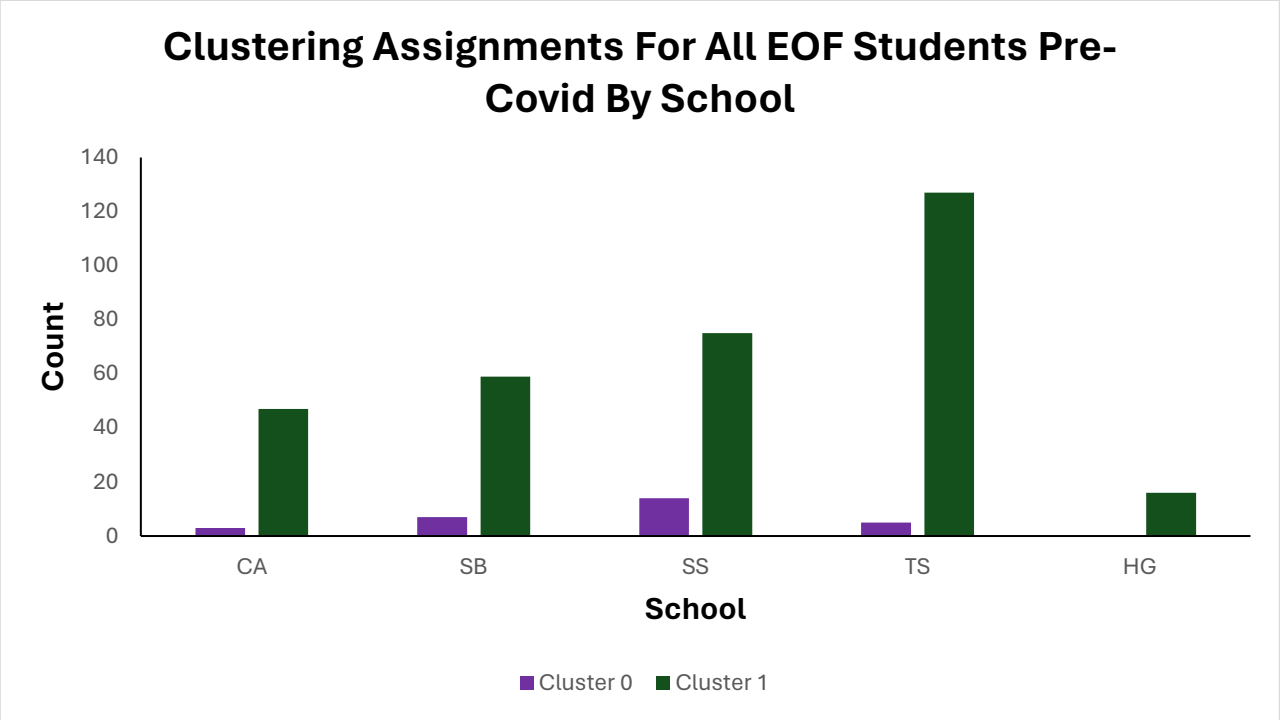


Figure 6.2.3 Clustering Assignments For All EOF Students Pre-Covid By School

Examining the residency status of the students within each cluster, cluster 1 has the largest number of commuters, 28 students, whereas cluster 0 has the least number of commuters, 18 students. Cluster 1 also has the highest number of residents, 342, and cluster 1 has the least number of residents, 13. This is depicted in Figure 6.2.4.

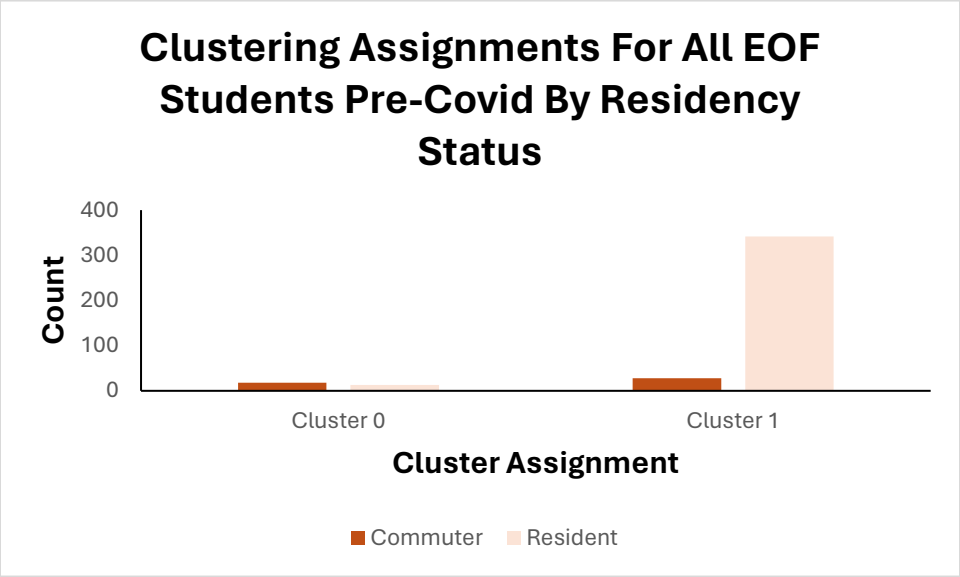


Figure 6.2.4 Clustering Assignments For All EOF Students Pre-Covid By Residency Status

When considering the academic performance within each cluster, which is measured by average cumulative GPA, there is a 0.10 grade point difference between students in cluster 0 and students in cluster 1, as cluster 0 has the higher average cumulative GPA.

Table 6.2.1 Average Cumulative GPA Per Cluster For All EOF Students Pre-Covid

Cluster Assignment	Average Cumulative GPA
0	2.906355
1	2.811692

Finally, Figure 6.2.5 shows retention rates amongst both clusters. Cluster 1 has the highest number of students not retaining ,34 or 9.2% of the cluster 1 population. Cluster 0 only has 1 student not retaining or 3.2% of the cluster 0 population not retaining.

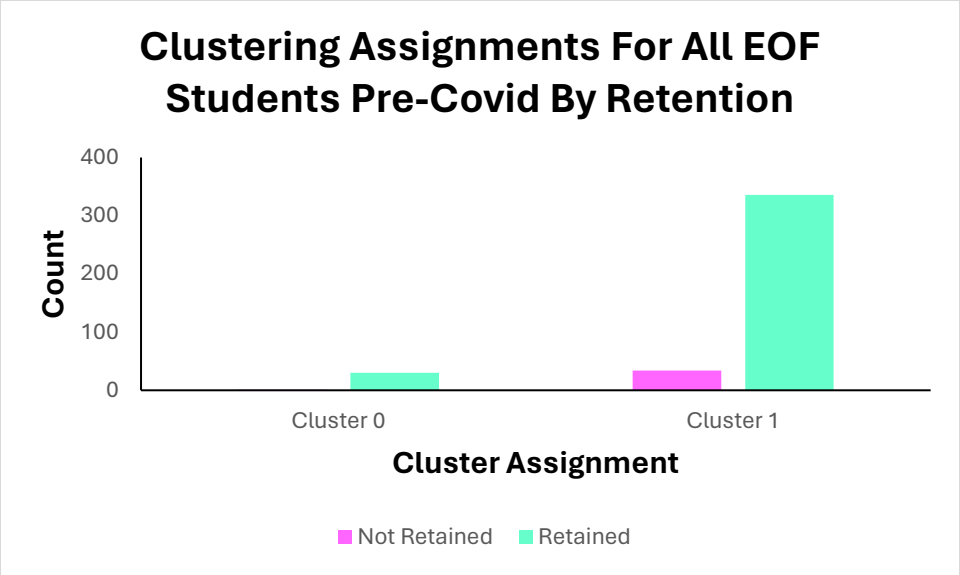


Figure 6.2.5 Clustering Assignments For All EOF Students By Retention

Section 6.3 Clustering All EOF Students Post-Covid

By iterating through the associated silhouette scores for the number of clusters within the range 2 through 10, the silhouette score for 2 clusters was the highest with a value of 0.56, so the EOF population post-covid was sorted into group 0 or group 1. This clustering, which is sparser due to the limited data compared to the previous two sections, along with each group's respective centroid is shown in Figure 6.3.1.

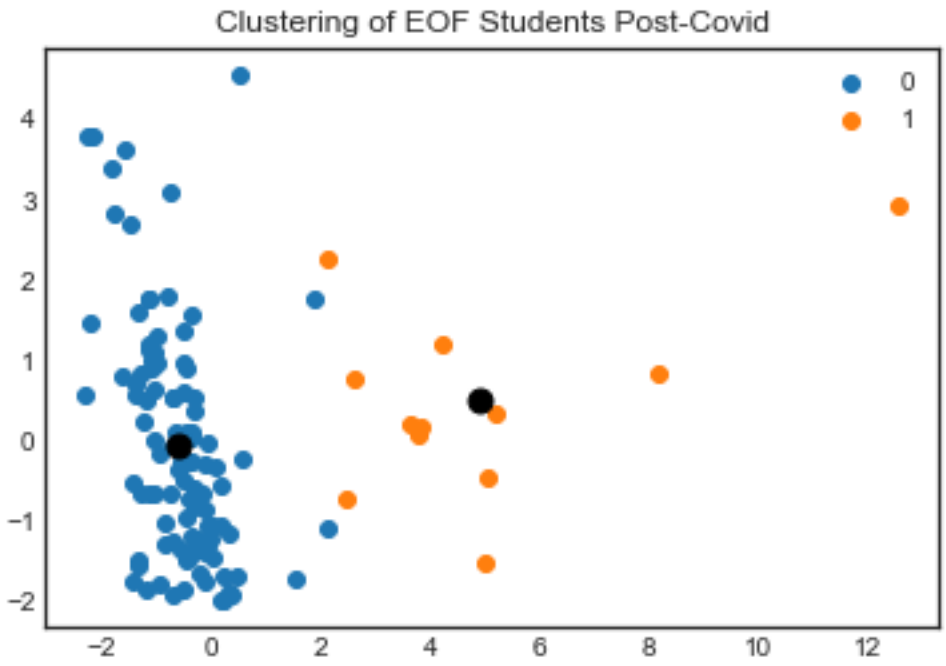
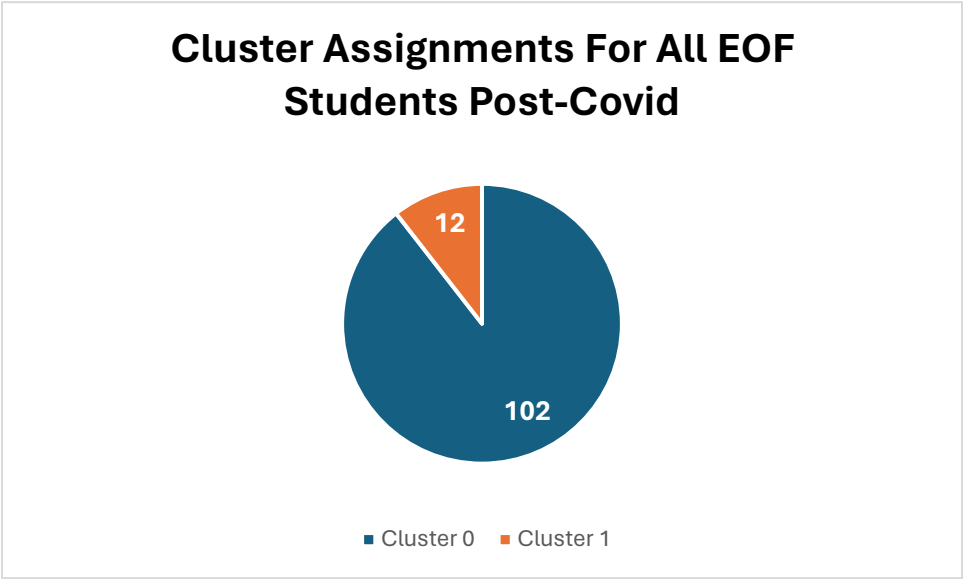


Figure 6.3.1 Clustering Of EOF Students Post-Covid

Figure 6.3.2 shows the disproportionate distribution of students within each group, where cluster 0 contains the majority of students, 102, followed by cluster 1 with 12 students.



Exploring the distribution of EOF students post-covid by school, cluster 0 contains the largest number of students from all five schools, Anisfield School of Business (SB), School of Social Science and Human Services (SS), School of Contemporary Arts (CA), School of Theoretical and Applied Science (TS), and School of Humanities and Global Studies (HG), with 20, 18, 10, 48, and 2 students, whereas cluster 1 has 2, 4, 2, 3, and 1 students within each school, respectively. These results are shown in Figure 6.3.3.

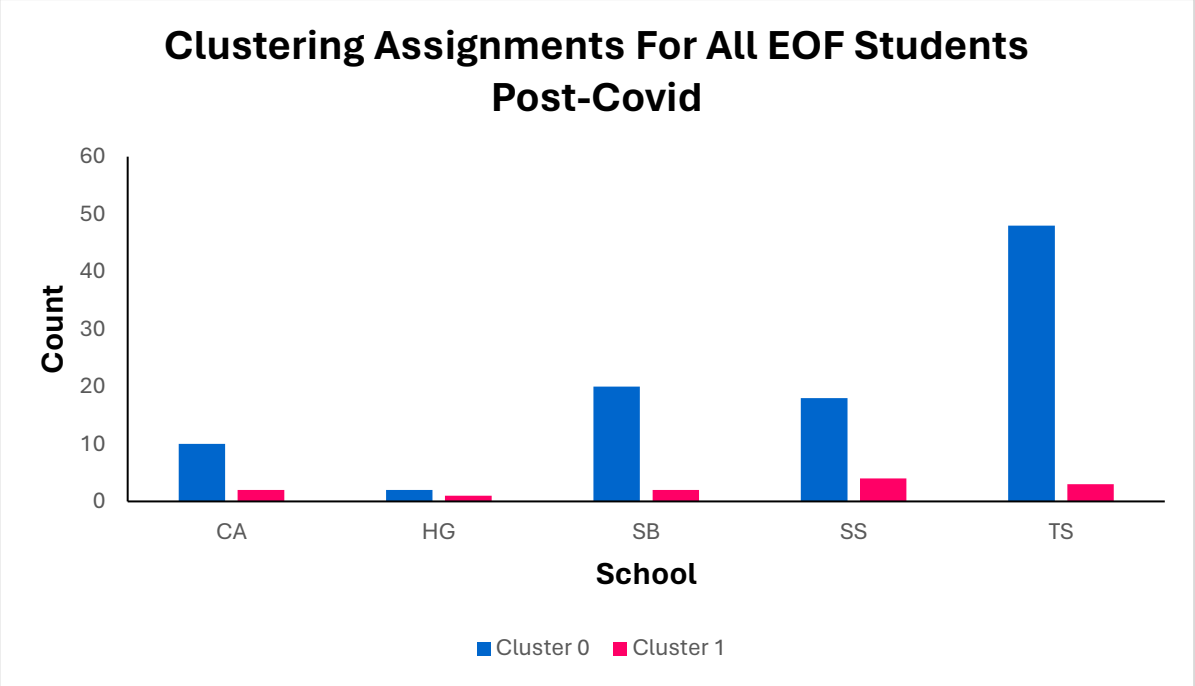


Figure 6.3.3 Clustering Assignments For All EOF Students Post-Covid By School

Examining residency status, cluster 0 has the highest number of students who are residents, 46, and cluster 1 contains the least number of on-campus residents, 4 students. Similarly, cluster 0 also has the highest composition of commuters, 56 students, and cluster 1 has the least number of commuters, 8. This is shown in Figure 6.3.4.

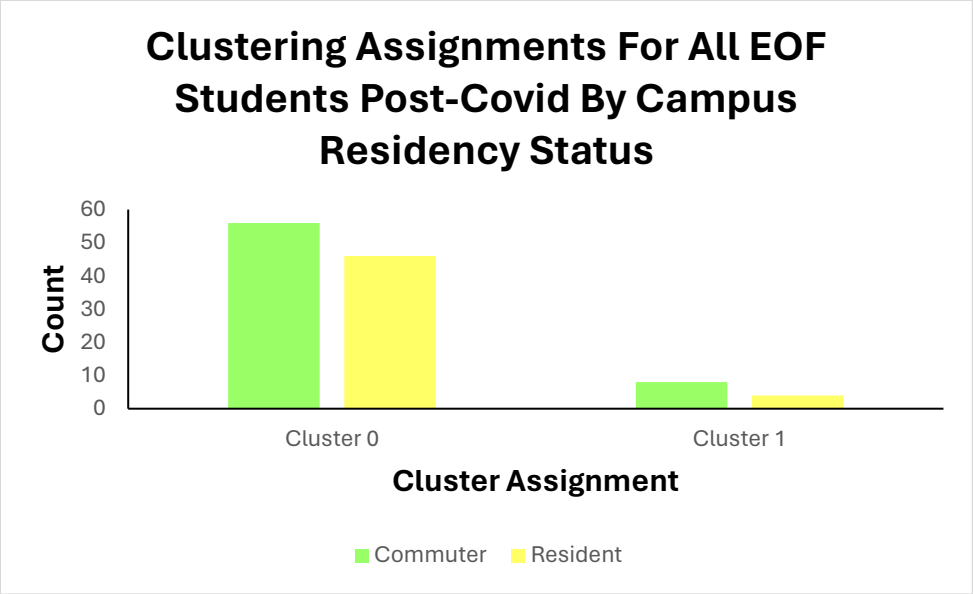


Figure 6.3.4 Clustering Assignments For All EOF Students Post-Covid By Residency Status

When considering the academic performance within each cluster, which is measured by average cumulative GPA, there is a 0.31 grade point difference between students in cluster 0 and students in cluster 1, as cluster 1 has the higher average cumulative GPA.

Table 6.3.1 Average Cumulative GPA Per Cluster For All EOF Students Post-Covid

Cluster Assignment	Average Cumulative GPA
0	2.750304
1	3.060333

Finally, Figure 6.3.5 shows retention rates amongst both clusters. Cluster 0 has the highest number of students not retaining ,21 or 20.6% of the cluster 0 population. Cluster 1 only has 4 students not retaining or 33% of the cluster 1 population not retaining.

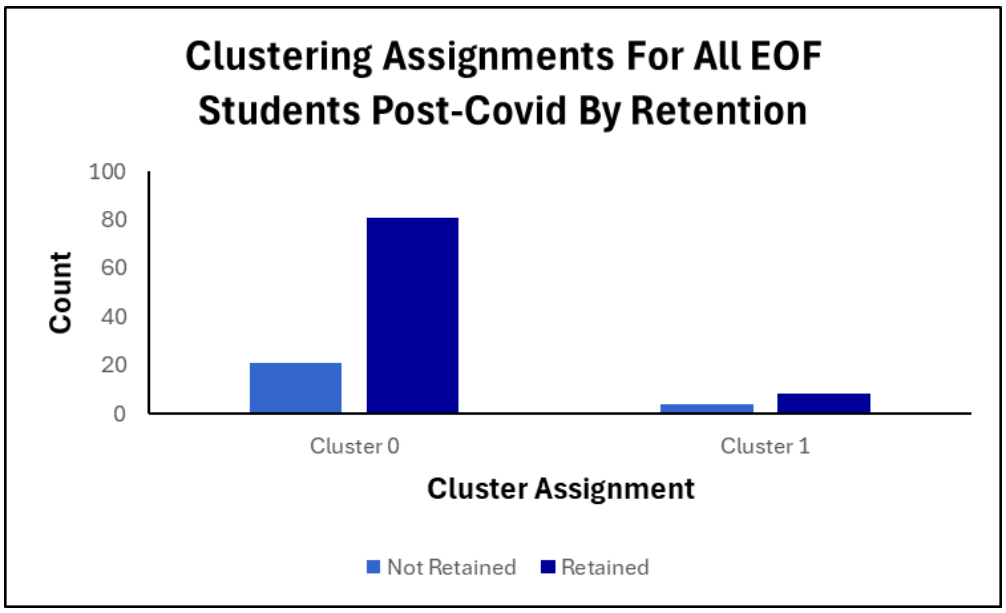


Figure 6.3.5 Clustering Assignments For All EOF Students Post-Covid By Retention

Chapter 7 Logistic Regression Results

Within this chapter, the logistic regression algorithm is used to predict whether or not an EOF student will retain. As previously stated, the analysis is broken down into 3 sections, all EOF students, all EOF students pre-covid, and all EOF students post-covid. Within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was 0 to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used and 75% of the data was used to train the model and 25% for testing, and the random state was set equal to 42 for reproducibility. The results for each implementation of the algorithm are shown in Table 7.1.1, where NR indicates 'Not Retained,' R indicates 'Retained,' and FS indicates 'Feature Selection.'

Since the focus of this analysis is on EOF students who are not retaining, the metric that

Table 7.1 Logistic Regression Model Performance

Logistic Regression Model	Precision NR	Recall NR	$F1$-Score NR	Support NR	Precision R	Recall R	$F1$-Score R	Support R
All EOF	0.00	0.00	0.00	12	0.91	1.00	0.95	117
EOF SMOTE	0.68	0.69	0.68	116	0.67	0.66	0.67	112
EOF FS	0.50	0.25	0.33	12	0.93	0.97	0.95	117
EOF FS & SMOTE	0.76	0.64	0.69	116	0.68	0.79	0.73	112
PRE-COVID	0.00	0.00	0.00	5	0.95	1.00	0.97	96
PRE-COVID SMOTE	0.68	0.62	0.65	91	0.65	0.72	0.68	92
PRE-COVID FS	0.33	0.20	0.25	5	0.96	0.98	0.97	96
PRE-COVID FS & SMOTE	0.74	0.59	0.66	91	0.66	0.79	0.72	92
POST-COVID	0.00	0.00	0.00	9	0.69	1.00	0.82	20
POST-COVID SMOTE	0.57	0.77	0.65	22	0.67	0.43	0.53	23
POST-COVID FS	0.00	0.00	0.00	9	0.69	1.00	0.82	20
POST-COVID FS & SMOTE	0.60	0.68	0.64	22	0.65	0.57	0.60	23

is most pertinent is the *Precision NR*. This metric details the accuracy of the students who were

predicted to not retain, specifically, out of the students who were predicted to not retain, how many of these predictions were correct. All of the models that implemented SMOTE had the highest precision scores, however the recommended model is the *EOF FS & SMOTE* which has a precision score of 0.76. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Regarding feature selection, Figure 7.1 shows the SHAP values that were used to determine the most important features, using a threshold of 0.5, for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

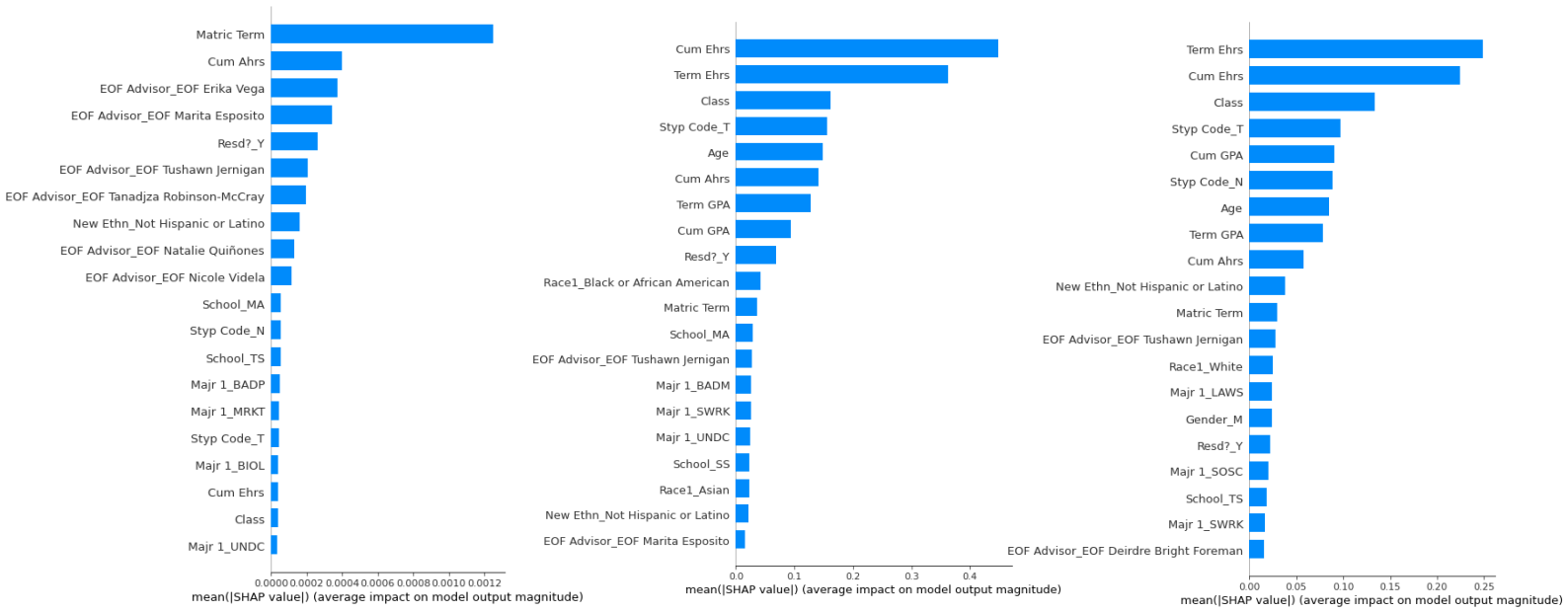


Figure 7.1 Logistic Regression Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

Specifically, for all EOF students the features were reduced to matric term, cumulative attempted hours (Cum Ahrs), EOF Advisor Erika Vega, EOF Advisor Marita Esposito, and campus resident (Resd?_Y). For all EOF students pre-covid, students the features were reduced to cumulative earned hours (Cum Ehrs), term earned hours (Term Ehrs), class, transfer student (Styp Code_T), cumulative attempted hours (Cum Ahrs), age, and term GPA. For all students post-covid, the features were reduced to term earned hours (Term Ehrs), cumulative earned hours (Cum Ehrs), class, transfer student (Styp Code_T), and age. Both the pre-covid and post-covid SHAP plots found the features term earned hours and cumulative earned hours to be significant.

Based on the results in Table 7.2, after performing 10-fold cross validation the models that implemented feature selection typically had the highest accuracy and precision, specifically the models *EOF FS,* and *Pre-Covid FS* out of their respective division. However, within the post-covid division, the *Post-Covid* model had the highest precision and accuracy. Therefore, Table 7.2 shows that the models *EOF FS, Pre-Covid FS,* and *Post-Covid* were the most accurate at predicting the retention of students.

In order to assess and compare the 10-fold cross validation results as seen in Table 7.2 to the logistic regression models in Table 7.1, the metric *Precision R* will be used. Within Table 7.2, the accuracy, precision, f-1, and recall scores are considered simultaneously for retention, whereas for the results in Table 7.1, these metrics are considered respectively for students who did retain (R), and for students who did not (NR). By focusing on the *Precision R* metric, which is the number of students who actually did retain out of the students predicted to, which is the same result at the *Precision* metric in Table 7.2, the models within each cohort that had highest and closest precision scores were *EOF FS, Pre-Covid FS,* and *Post Covid FS.* Out of all three of these models, the recommended model is *EOF FS,* as even though this model did not have the

highest precision scores, *Precision (R)* is 0.93 and the 10-fold cross validation *Precision* is 0.94, these metrics were obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Table 7.2 Logistic Regression Model Evaluation With 10-fold Cross Validation

Logistic Regression Model	Accuracy	Precision	$F1$-Score	Recall
All EOF	0.88	0.88	0.93	1.0
EOF SMOTE	0.71	0.66	0.66	0.66
EOF FS	0.89	0.90	0.94	0.98
EOF FS & SMOTE	0.75	0.74	0.76	0.79
PRE-COVID	0.9	0.9	0.94	0.99
PRE-COVID SMOTE	0.74	0.72	0.76	0.75
PRE-COVID FS	0.91	0.92	0.95	0.99
PRE-COVID FS & SMOTE	0.74	0.71	0.75	0.82
POST-COVID	0.82	0.82	0.89	0.98
POST-COVID SMOTE	0.70	0.78	0.6	0.5
POST-COVID FS	0.79	0.81	0.88	0.97
POST-COVID FS & SMOTE	0.71	0.74	0.65	0.6

Chapter 8 Decision Tree Classifier Results

Within this chapter, the decision tree classifier is used to predict whether an EOF student will retain, where the maximum depth is set to 5, and the random state is set equal to 42 for reproducibility. As previously stated, the analysis is broken down into 3 sections, all EOF students, all EOF students pre-covid, and all EOF students post-covid. Within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was 0 to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used and 75% of the data was used to train the model and 25% for testing, and the random state was set equal to 42 for reproducibility. The results for each implementation of the algorithm are shown in Table 8.1, where NR indicates ‘Not Retained’, R indicates ‘Retained’, and FS indicates ‘Feature Selection’.

Table 8.1 Decision Tree Model Performance

Decision Tree Classifier Model	Precision NR	Recall NR	*F*1-Score NR	Support NR	Precision R	Recall R	*F*1-Score R	Support R
All EOF	0.42	0.42	0.42	12	0.94	0.94	0.94	117
EOF SMOTE	0.81	0.87	0.84	116	0.86	0.79	0.82	112
EOF FS	0.38	0.42	0.40	12	0.94	0.93	0.94	117
EOF FS & SMOTE	0.84	0.76	0.80	116	0.77	0.85	0.81	112
PRE-COVID	0.40	0.40	0.40	5	0.97	0.97	0.97	96
PRE-COVID SMOTE	0.88	0.57	0.69	91	0.69	0.92	0.79	92
PRE-COVID FS	0.40	0.40	0.40	5	0.97	0.97	0.97	96
PRE-COVID FS & SMOTE	0.88	0.57	0.69	91	0.69	0.92	0.79	92
POST-COVID	0.50	0.33	0.40	9	0.74	0.85	0.79	20
POST-COVID SMOTE	0.94	0.77	0.85	22	0.81	0.96	0.88	23
POST-COVID FS	0.80	0.44	0.57	9	0.79	0.95	.86	20
POST-COVID FS & SMOTE	0.93	0.64	0.76	22	0.73	0.96	0.83	235

Similarly to the process used in logistic regression, the models in Table 8.1 were evaluated using the *Precision NR* as the metric. All the models that implemented both SMOTE and feature selection had the highest precision scores, however the recommended model is the *EOF FS & SMOTE* which has a precision score of 0.84. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Regarding feature selection, Figure 8. Shows the summary plot of the SHAP values that were used to determine the most important features for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

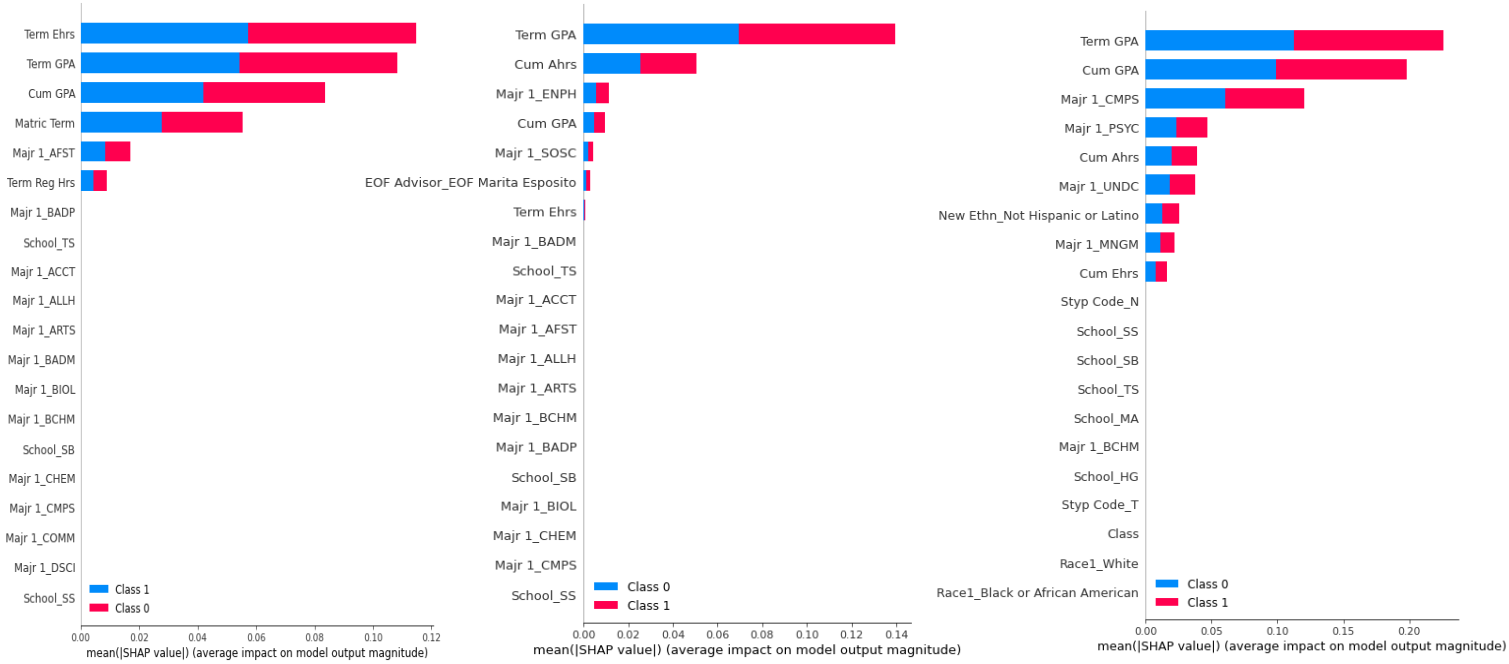


Figure 8.1 Decision Tree Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

For all EOF students the features were reduced to term earned hours (Term Ehrs), term GPA, cumulative GPA, and matric term. For all EOF students pre-covid, students the features were reduced term GPA, and cumulative attempted hours (Cum Ahrs). Finally, for all students post-covid, the features were reduced to term GPA, and cumulative GPA. All three analyses found the predictor term GPA to be significant, whereas only the post-covid and all EOF student analyses found the predictor cumulative GPA to be significant.

Based on the results in Table 8.2, after performing 10-fold cross validation the models that implemented feature selection typically had the highest accuracy and precision, specifically, the models *EOF FS*, *Pre-Covid FS*, and *Post-Covid*, out of their respective division.

In order to assess and compare the 10-fold cross validation results as seen in Table 8.2 to the decision tree models in Table 8.1, the metric *Precision R* will be used. Within Table 8.2, the accuracy, precision, f-1, and recall scores are considered simultaneously for retention, whereas

for the results in Table 8.1, these metrics are considered respectively for students who did retain (R), and for students who did not (NR). By focusing on the *Precision R* metric, which is the number of students who actually did retain out of the students predicted to, which is the same result at the *Precision* metric in Table 8.2, the models within each cohort that had highest and closest precision scores were *EOF SMOTE*, *Pre-Covid FS*, and *Post Covid SMOTE*. Out of all three of these models, the recommended model is *EOF SMOTE*, as even though this model did not have the highest precision scores, *Precision (R)* is 0.86 and the 10-fold cross validation *Precision* is 0.88, these metrics were obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Table 8.2 Decision Tree Model Evaluation With 10-fold Cross Validation

Decision Tree Classifier Model	Accuracy	Precision	$F1$-Score	Recall
All EOF	0.87	0.90	0.93	0.96
EOF SMOTE	0.86	0.88	0.86	0.85
EOF FS	0.88	0.91	0.93	0.96
EOF FS & SMOTE	0.78	0.79	0.77	0.76
PRE-COVID	0.89	0.92	0.94	0.96
PRE-COVID SMOTE	0.80	0.74	0.82	0.92
PRE-COVID FS	0.9	0.93	0.94	0.96
PRE-COVID FS & SMOTE	0.77	0.71	0.8	0.92
POST-COVID	0.81	0.89	0.87	0.87
POST-COVID SMOTE	0.81	0.83	0.78	0.74
POST-COVID FS	0.88	0.93	0.92	0.91
POST-COVID FS & SMOTE	0.74	0.7	0.72	0.76

Chapter 9 Random Forest Classifier Results

Within this chapter, random forest classifier models are used to predict whether or not an EOF student will retain, and just like the previous chapter discussing the decision tree results, the maximum depth is set to 5, and the random state is set equal to 42 for reproducibility. Like the other chapters, the analysis is broken down into 3 sections, all EOF students, all EOF students pre-covid, and all EOF students post-covid. Within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was 0 to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used and 75% of the data was used to train the model and 25% for testing, and the random state was set equal to 42 for reproducibility. The results for each implementation of the algorithm are shown, where NR indicates 'Not Retained,' R indicates 'Retained,' and FS indicates 'Feature Selection.'

Table 9.1 Random Forest Classifier Performance

Random Forest Classifier Model	Precision NR	Recall NR	$F1$-Score NR	Support NR	Precision R	Recall R	$F1$-Score R	Support R
All EOF	0.43	0.25	0.32	12	0.93	0.97	0.95	117
EOF SMOTE	0.90	0.82	0.86	116	0.83	0.91	0.87	112
EOF FS	0.44	0.33	0.38	12	0.93	0.96	0.95	117
EOF FS & SMOTE	0.96	0.44	0.60	116	0.63	0.98	0.77	112
PRE-COVID	0.25	0.20	0.22	5	0.96	0.97	0.96	96
PRE-COVID SMOTE	0.94	0.32	0.48	91	0.59	0.98	0.74	92
PRE-COVID FS	0.20	0.20	0.20	5	0.96	0.96	0.96	96
PRE-COVID FS & SMOTE	0.91	0.85	0.87	91	0.86	0.91	0.88	92
POST-COVID	0.67	0.22	0.33	9	0.73	0.95	0.83	20
POST-COVID SMOTE	0.90	0.82	0.86	22	0.84	0.91	0.87	23
POST-COVID FS	0.67	0.22	0.33	9	0.73	0.95	0.83	20
POST-COVID FS & SMOTE	1.00	0.45	0.62	22	0.66	1.00	0.79	23

Using the *Precision NR* metric to compare the models, the models that implemented both SMOTE and feature selection had the highest precision scores within their respective category of all EOF students, pre-covid, and post-covid, respectively. However, the recommended model is the *EOF FS & SMOTE* which has a precision score of 0.96. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to just the pre-covid and post-covid subsections.

Regarding feature selection, Figure 9.1 shows the summary plot of the SHAP values that were used to determine the most important features, for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

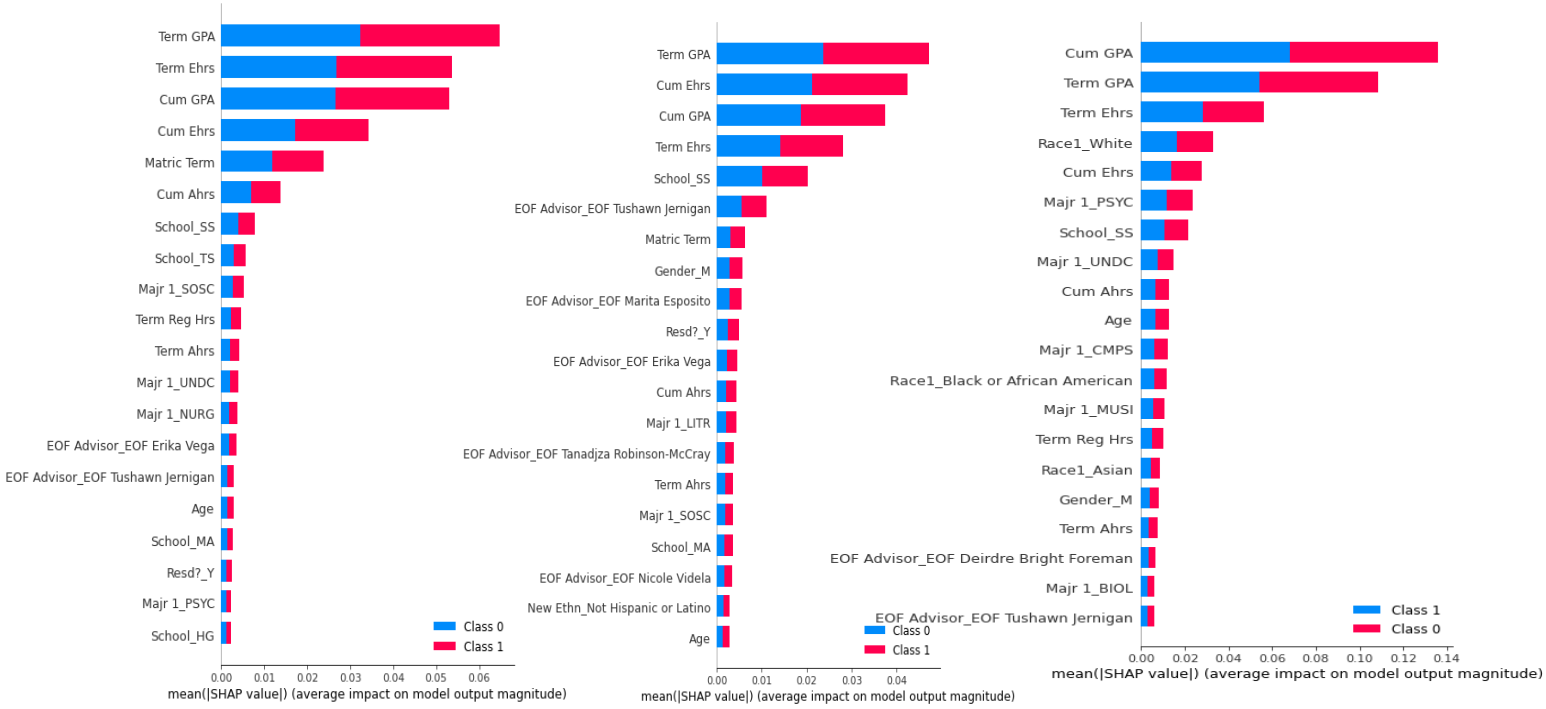


Figure 9.1 Random Forest Classifier Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

As shown in Figure 9.1, for all EOF students the features were reduced to term GPA, term earned hours (Term Ehrs), cumulative GPA, cumulative earned hours (Cum Ehrs), and matric term. For all EOF students pre-covid, the features were reduced to term GPA, cumulative earned hours (Cum Ehrs), cumulative GPA, term earned hours (Term Ehrs) and the School of Social Science and Human Services. Finally, for all students post-covid, the features were reduced cumulative GPA, term GPA, and term earned hours (Term Ehrs). All three analyses found the features term GPA, cumulative GPA (Cum GPA), and term earner hours (Term Ehrs) significant.

In order to rigorously evaluate how each model performed, 10-fold cross validation was performed, and the results are shown in Table 9.2. Unlike the results in previous chapters, the models that had the highest precision and accuracy were *All EOF, EOF FS, Pre-Covid,* and *Post-Covid.* Interestingly, out of the four models previously mentioned, three of them were imbalanced and did not utilize feature selection.

In order to assess and compare the 10-fold cross validation results as seen in Table 9.2 to the random forest models in Table 9.1, the metric *Precision R* will be used. Within Table 9.2, the accuracy, precision, f-1, and recall scores are considered simultaneously for retention, whereas for the results in Table 9.1, these metrics are considered respectively for students who did retain (R), and for students who did not (NR). By focusing on the *Precision R* metric, which is the number of students who actually did retain out of the students predicted to, which is the same result at the *Precision* metric in Table 9.2, the models within each cohort that had highest and closest precision scores were *EOF FS, Pre-Covid, Pre-Covid FS, Post-Covid,* and *Post-Covid FS*. Out of all three of these models, the recommended model is *EOF FS,* as even though this model did not have the highest precision scores, *Precision (R)* is 0.93 and the 10-fold cross

validation *Precision* is 0.91, these metrics were obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Table 9.2 Random Forest Classifier Evaluation With 10-fold Cross Validation

Random Forest Classifier Model	Accuracy	Precision	$F1$-Score	Recall
All EOF	0.89	0.9	0.94	0.98
EOF SMOTE	0.88	0.86	0.88	0.91
EOF FS	0.89	0.91	0.94	0.97
EOF FS & SMOTE	0.82	0.79	0.83	0.87
PRE-COVID	0.91	0.92	0.95	0.99
PRE-COVID SMOTE	0.91	0.91	0.91	0.91
PRE-COVID FS	0.9	0.92	0.95	0.97
PRE-COVID FS & SMOTE	0.86	0.85	0.86	0.87
POST-COVID	0.86	0.88	0.91	0.95
POST-COVID SMOTE	0.82	0.8	0.8	0.81
POST-COVID FS	0.6	0.9	0.91	0.92
POST-COVID FS & SMOTE	0.77	0.76	0.75	0.77

Chapter 10 Gradient Boosting Classifier Results

Within this chapter gradient boosting classifier models are used to predict whether or not an EOF student will retain, where the learning state is 0.1, and the random state is set equal to 42 for reproducibility. Like the preceding chapters, the analysis is broken down into 3 sections, all EOF students, all EOF students pre-covid, and all EOF students post-covid. Then within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was 0 to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used and 75% of the data was used to train the model and 25% for testing, and the random state was set equal to 42, for reproducibility. The results for each implementation of the algorithm are shown in Table 10.1, where NR indicates 'Not Retained,' R indicates 'Retained,' and FS indicates 'Feature Selection.'

Table 10.1 Gradient Boosting Classifier Performance

Gradient Boosting Classifier Model	Precision NR	Recall NR	$F1$-Score NR	Support NR	Precision R	Recall R	$F1$-Score R	Support R
All EOF	0.36	0.42	0.0.38	12	0.94	0.92	0.93	117
EOF SMOTE	0.91	0.92	0.91	116	0.92	0.90	0.91	112
EOF FS	0.33	0.42	0.37	12	0.94	0.91	0.93	117
EOF FS & SMOTE	0.87	0.83	0.85	116	0.83	0.88	0.85	112
PRE-COVID	0.20	0.20	0.20	5	0.96	0.96	0.96	96
PRE-COVID SMOTE	0.93	0.86	0.89	91	0.87	0.93	0.90	92
PRE-COVID FS	0.33	0.40	0.36	5	0.97	0.96	0.96	96
PRE-COVID FS & SMOTE	0.91	0.85	0.87	91	0.86	0.91	0.88	92
POST-COVID	0.57	0.44	0.50	9	0.77	0.85	0.81	20
POST-COVID SMOTE	0.90	0.86	0.88	22	0.88	0.91	0.89	23
POST-COVID FS	0.43	0.33	0.38	9	0.73	0.80	0.76	20
POST-COVID FS & SMOTE	0.77	0.77	0.77	22	0.78	0.78	0.78	23

As shown in Table 10.1, using the *Precision NR* metric to compare the models, the models that implemented SMOTE, specifically *EOF SMOTE, PRE-COVID SMOTE,* and *POST-COVID SMOTE,* had the highest precision scores within their respective categories. However, the recommended model is *EOF SMOTE* which has a precision score of 0.91. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to just the pre-covid and post-covid subsections.

Regarding feature selection, Figure 10.1 shows the summary plot of the SHAP values that were used to determine the most important features, for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

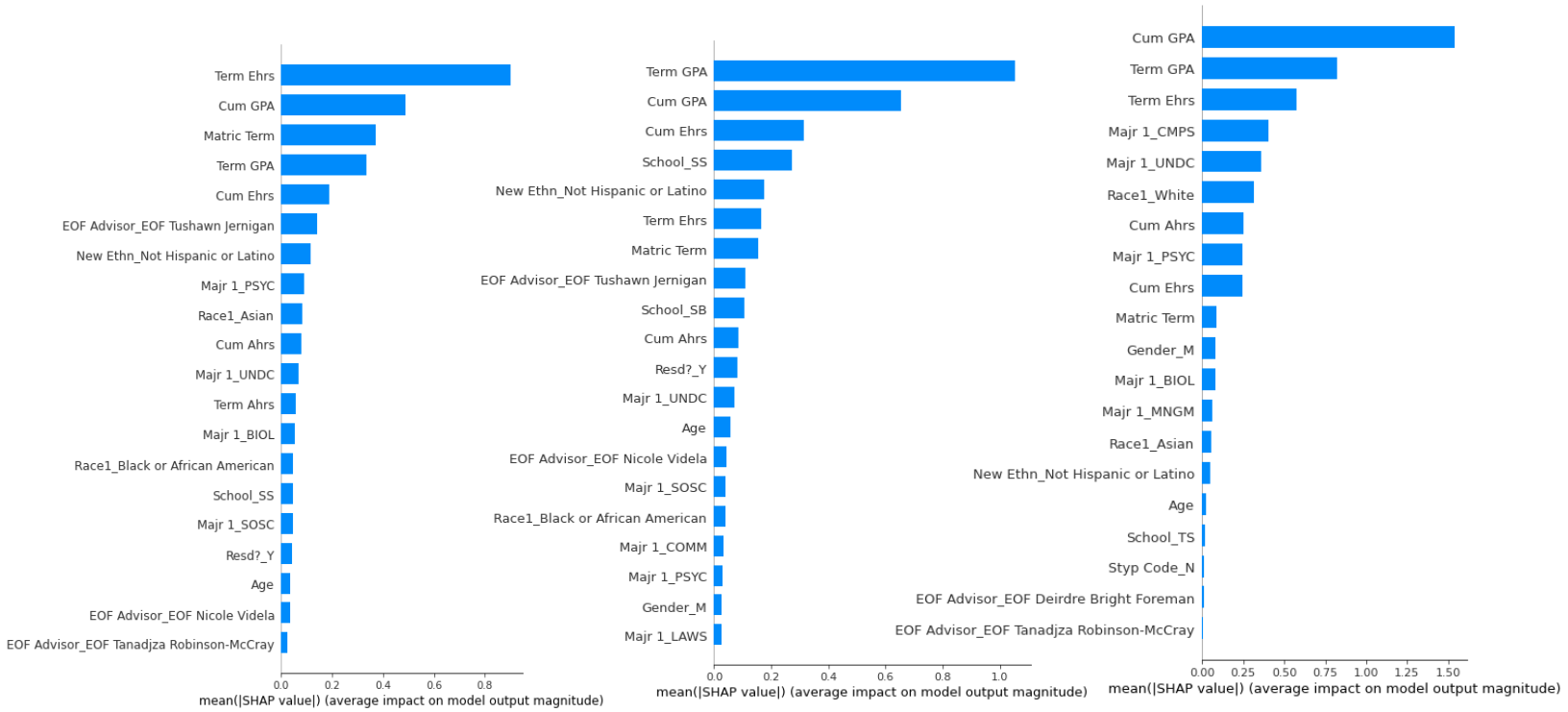


Figure 10.1 Gradient Boosting Classifier Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

As seen from Figure 10.1, for all EOF students the features were reduced to term earned hours (Term Ehrs), cumulative GPA, matric term, term GPA, and cumulative earned hours (Cum Ehrs). For all EOF students pre-covid, the features were reduced to term GPA, cumulative GPA, cumulative earned hours (Cum Ehrs), the School of Social Science and Human Services, and term earned hours (Term Ehrs). Finally, for all students post-covid, the features were reduced to cumulative GPA, term GPA, term earned hours (Term Ehrs), computer science major (Majr 1_CMPS), and undeclared major (Majr 1_UNDC). All three analyses found the features term GPA, cumulative GPA (Cum GPA), and term earned hours (Term Ehrs) significant.

In order to rigorously evaluate how each model performed, 10-fold cross validation was performed, and the results are shown in Table 10.2. Within each section the models that had the highest accuracy and precision scores were *EOF SMOTE*, *PRE-COVID SMOTE*, and *POST-COVID FS*. For all EOF students and all EOF students pre-covid, both of these models implemented SMOTE, however for post-covid this model was imbalanced, but feature selection was utilized.

In order to assess and compare the 10-fold cross validation results as seen in Table 10.2 to the random forest models in Table 10.1, the metric *Precision R* will be used. Within Table 10.2, the accuracy, precision, f-1, and recall scores are considered simultaneously for retention, whereas for the results in Table 10.1, these metrics are considered respectively for students who did retain (R), and for students who did not (NR). By focusing on the *Precision R* metric, which is the number of students who actually did retain out of the students predicted to, which is the same result at the *Precision* metric in Table 10.2, the models within each cohort that had highest and closest precision scores were *EOF FS*, *Pre-Covid*, *Pre-Covid FS*, and *Post-Covid SMOTE*. Out of all three of these models, the recommended model is *EOF FS*, as even though this model

did not have the highest precision scores, *Precision (R)* is 0.94 and the 10-fold cross validation *Precision* is 0.92, these metrics were obtained using all of the data, as opposed to the pre-covid and post-covid subsections.

Table 10.2 Gradient Boosting Classifier Evaluation With 10-fold Cross-Validation

Gradient Boosting Classifier Model	Accuracy	Precision	$F1$-Score	Recall
All EOF	0.89	0.91	0.94	0.97
EOF SMOTE	0.94	0.93	0.93	0.94
EOF FS	0.89	0.92	0.94	0.96
EOF FS & SMOTE	0.86	0.87	0.86	0.85
PRE-COVID	0.91	0.93	0.95	0.97
PRE-COVID SMOTE	0.95	0.97	0.95	0.93
PRE-COVID FS	0.9	0.93	0.95	0.97
PRE-COVID FS & SMOTE	0.86	0.86	0.86	0.86
POST-COVID	0.83	0.91	0.89	0.9
POST-COVID SMOTE	0.84	0.82	0.81	0.8
POST-COVID FS	0.87	0.92	0.92	0.93
POST-COVID FS & SMOTE	0.77	0.77	0.75	0.75

Chapter 11 Support Vector Machine Results

Within this chapter support vector machine models are used to predict whether or not an EOF student will retain, where a linear kernel is used, and the random state is set equal to 42 for reproducibility. Like the preceding chapters, the analysis is broken down into 3 sections, all EOF students, all EOF students pre-covid, and all EOF students post-covid. Then within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was 0 to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used, which ensured that the algorithm could understand the variables it was being fed, and 75%-25% training, testing split was used. The results for each implementation of the algorithm are shown in Table 11.1, where NR indicates 'Not Retained,' R indicates 'Retained,' and FS indicates 'Feature Selection.'

Table 11.1 Support Vector Machine Model Performance

Support Vector Machine Model	Precision NR	Recall NR	$F1$-Score NR	Support NR	Precision R	Recall R	$F1$-Score R	Support R
All EOF	0.00	0.00	0.00	12	0.91	1.00	0.95	117
EOF SMOTE	0.57	0.80	0.67	116	0.65	0.38	0.48	112
EOF FS	0.00	0.00	0.00	12	0.91	1.00	0.95	117
EOF FS & SMOTE	0.59	0.66	0.62	116	0.60	0.54	0.57	112
PRE-COVID	0.00	0.00	0.00	5	0.92	0.64	0.75	96
PRE-COVID SMOTE	0.55	0.97	0.70	91	0.87	0.22	0.35	92
PRE-COVID FS	0.11	0.40	0.17	5	0.96	0.82	0.89	96
PRE-COVID FS & SMOTE	0.54	0.48	0.51	91	0.54	0.60	0.57	92
POST-COVID	1.00	0.11	0.20	9	0.71	1.00	0.83	20
POST-COVID SMOTE	1.00	0.18	0.31	22	0.56	1.00	0.72	23
POST-COVID FS	0.33	0.11	0.17	9	0.69	0.90	0.78	20
POST-COVID FS & SMOTE	0.68	0.59	0.63	22	0.65	0.74	0.69	23

Using the *Precision NR* metric to compare the models, the best models in each category were *EOF FS & SMOTE, PRE-COVID SMOTE,* and *POST-COVID SMOTE,* respectively. Each of these models did implement SMOTE, and the overall recommended model is the *EOF FS & SMOTE* which has a precision score of 0.59. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to just the pre-covid and post-covid subsections.

Regarding feature selection, Figure 11.1 shows the summary plot of the SHAP values that were used to determine the most important features, for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

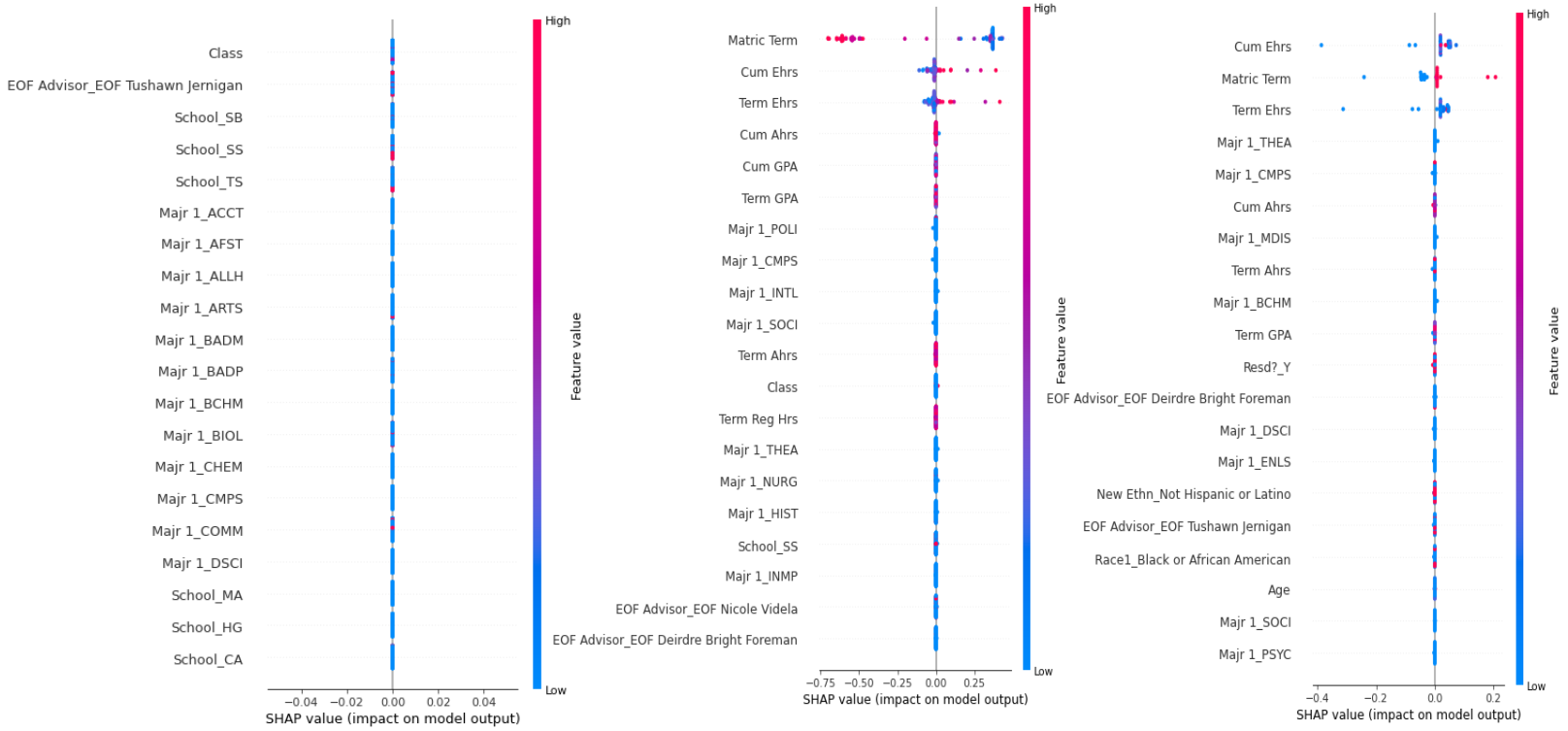


Figure 11.1 Support Vector Machine Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

As seen from Figure 11.1, for all EOF students the features were reduced to class, EOF advisor Tushawn Jernigan, Anisfield School of Business, School of Social Science and Human Services, and the School of Theoretical and Applied Science. For all EOF students pre-covid, the features were reduced to matric term, cumulative earned hours (Cum Ehrs), and term earned hours (Term Ehrs). Finally, for all students post-covid, the features were reduced to term earned hours (Term Ehrs), cumulative attempted hours (Cum Ahrs), matric term, and cumulative earned hours (Cum Ehrs). The pre-covid and post-covid analyses found the features matric term, term earned hours (Term Ehrs), and cumulative earned hours (Cum Ehrs) significant.

Unfortunately, 10-fold cross validation was not able to be performed for these models, as it was unable to ever produce an output. This is most likely a result of not having enough processing power.

Chapter 12 Ensemble Results

Within this chapter, an ensemble model comprised of the models, logistic regression, random forest, and support vector machine, use hard voting to predict whether or not an EOF student will retain, and the random state is set equal to 42 for reproducibility. Like the preceding chapters, the analysis is broken down into three sections: all EOF students, all EOF students pre-covid, and all EOF students post-covid. Then within each section, the analysis is divided further into predicting retention using the data, using the data with smote, using the data with feature selection, and using the data with feature selection and SMOTE.

For each implementation of the algorithm, the predictor was set to be the *Retention* column, which was zero to indicate that a student did not retain, and 1 to indicate that a student did. The encoded dataframe was used, which ensured that the algorithm could understand the variables it was being fed, and 75%-25% training, testing split was used. The results for each implementation of the algorithm are shown in Table 12.1, where NR indicates 'Not Retained,' R indicates 'Retained,' and FS indicates 'Feature Selection.'

Table 12.1 Ensemble Model Performance

Ensemble Model	Precision NR	Recall NR	$F1$-Score NR	Support NR	Precision R	Recall R	$F1$-Score R	Support R
All EOF	0.00	0.00	0.00	12	0.91	1.00	0.95	113
EOF SMOTE	0.69	0.72	0.71	116	0.70	0.66	0.68	112
EOF FS	0.00	0.00	0.00	12	0.91	1.00	0.95	117
EOF FS & SMOTE	0.60	0.66	0.63	116	0.61	0.54	0.57	112
PRE-COVID	0.00	0.00	0.00	5	0.95	1.00	0.97	96
PRE-COVID SMOTE	0.74	0.82	0.78	91	0.80	0.71	0.75	92
PRE-COVID FS	0.50	0.20	0.29	5	0.96	0.99	0.97	96
PRE-COVID FS & SMOTE	0.82	0.60	0.70	91	0.69	0.87	0.77	92
POST-COVID	0.00	0.00	0.00	9	0.69	1.00	0.82	20
POST-COVID SMOTE	0.89	0.73	0.80	22	0.78	0.91	0.84	23
POST-COVID FS	0.00	0.00	0.00	9	0.69	1.00	0.82	20
POST-COVID FS & SMOTE	1.00	0.64	0.78	22	0.74	1.00	0.85	23

Using the *Precision NR* metric to compare the models, the best models in each category were *EOF SMOTE, PRE-COVID Fs & SMOTE,* and *POST-COVID FS & SMOTE,* respectively. Each of these models did implement SMOTE, however the models for pre-covid and post-covid also utilized feature selection. The overall recommended model is the *EOF SMOTE* which has a precision score of 0.69. Even though this was not the highest out of all of the precision scores, this metric was obtained using all of the data, as opposed to just the pre-covid and post-covid subsections.

Regarding feature selection, Figure 12.1 shows the summary plot of the SHAP values that were used to determine the most important features, for all EOF students, EOF students pre-covid, and EOF students post-covid, respectively.

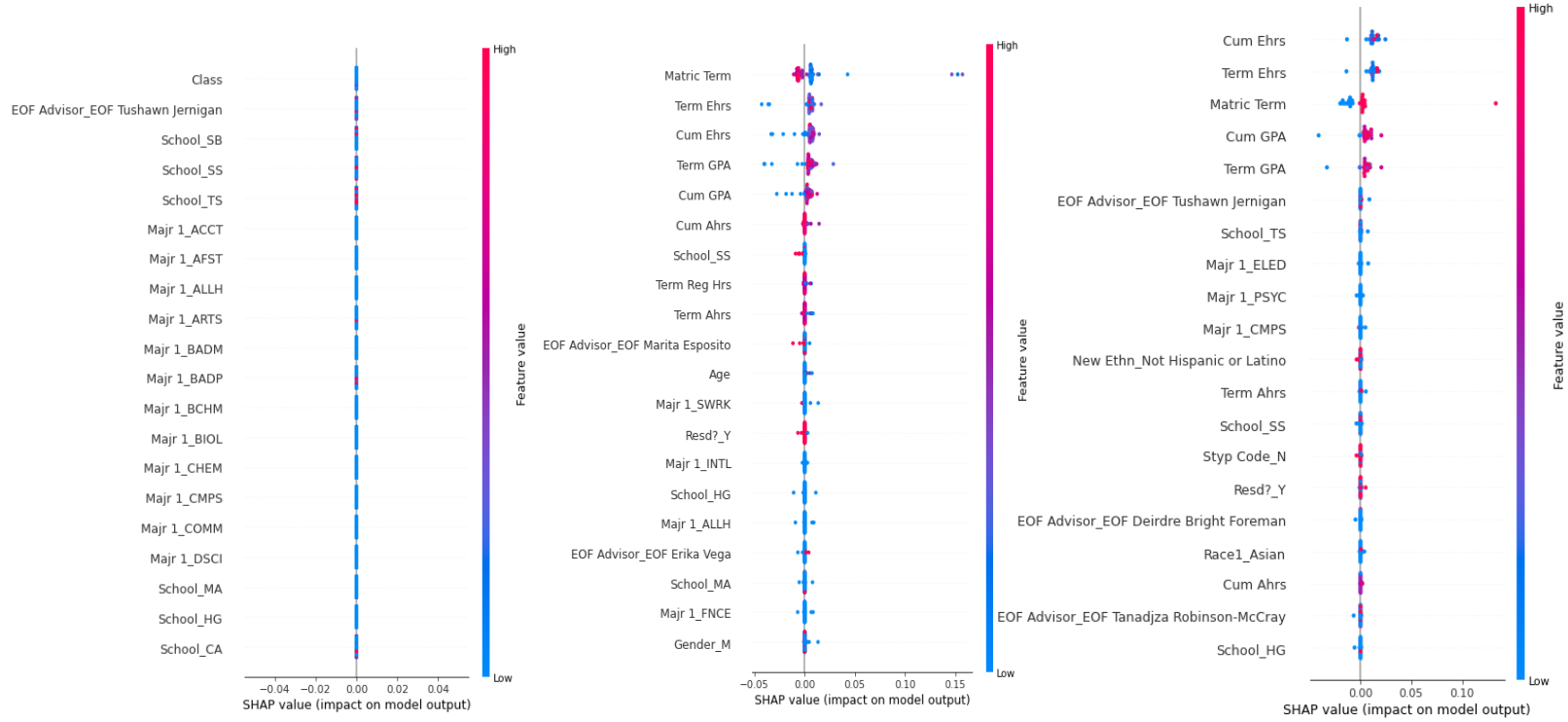


Figure 12.1 Ensemble Predictor SHAP Values For All EOF Students, EOF Students Pre-Covid, and EOF Students Post-Covid

As shown in Figure 12.1, the most important features for all EOF students were class, EOF advisor Tushawn Jernigan, the Anisfield School of Business, the School of Social Science

and Human Services, and the School of Theoretical and Applied Science. For EOF students pre-covid and post-covid, the most important features were matric term, term earned hours (Term Ehrs), and cumulative earned hours (Cum Ehrs). Unfortunately, 10-fold cross-validation was not able to be performed for these models, as it was unable to ever produce an output. This is most likely a result of not having enough processing power.

Chapter 13 Predicting First-Year EOF Retention

Discussion

For each method as discussed in Chapters 7 through 12, Table 13.1 provides a comparison of the recommended models and their associated metric, the precision score for EOF attrition. All of the recommended models implemented SMOTE, and two thirds of the recommended models implemented both SMOTE and feature selection. Out of all the models within Table 13.1, it is recommended that the EOF department use the Random Forest Classifier *EOF FS & SMOTE* model since it had the highest precision of 0.96 for predicting EOF student attrition.

Table 13.1 Comparing Recommended Model and Precision Scores For All Machine Learning Models Implemented

Method	**Recommended Model**	**Precision NR**
Logistic Regression	EOF FS & SMOTE	0.76
Decision Tree Classifier	EOF FS & SMOTE	0.84
Random Forest Classifier	EOF FS & SMOTE	0.96
Gradient Boosting Classifier	EOF SMOTE	0.91
Support Vector Machine	EOF FS & SMOTE	0.59
Ensemble	EOF SMOTE	0.69

Since two thirds of the recommended models in Table 13.1 implemented feature selection (FS), Table 13.2 compares the features that were found to be significant for the logistic regression, decision tree classifier, random forest classifier, and support vector machine methods.

Table 13.2 Comparing Recommended Models Implementing Feature Selection

Method	**Recommended Model**	**SHAP Feature Selection**
Logistic Regression	EOF FS & SMOTE	• Matric term • Cum attempted hours • EOF advisor Erika Vega • EOF advisor Marita Esposito • Campus Resident
Decision Tree Classifier	EOF FS & SMOTE	• Term earned hours • Term GPA • Cumulative GPA • Matric term
Random Forest Classifier	EOF FS & SMOTE	• Term GPA • Term earned hours • Cumulative GPA • Cumulative earned hours • Matric term
Support Vector Machine	EOF FS & SMOTE	• Class • EOF advisor Tushawn Jernigan • Anisfield School of Business • School of Social Science and Human Services • School of Theoretical and Applied Science

As shown in Table 13.2, for three out of the four methods, specifically logistic regression, decision tree classifier, and random forest classifier, all of their recommended models implementing feature selection found the features matric term to be significant. Unsurprisingly, given the relationship between the decision tree and random forest methods, they both found the features term earned hours (Term Ehrs), term GPA, and cumulative GPA significant.

Unfortunately, 10-fold cross validation was not able to be implemented for each method, but Table 13.3 below provides a comparison of the recommended models and their associated precision metric. Three out of the four recommended models below implement feature selection, suggesting the importance of performing this for future EOF retention studies.

Table 13.3 Comparing Recommended Models Implementing 10-Fold Cross Validation

Method	**Recommended Model**	**Precision (R)**	**Precision 10-Fold Cross Validation**
Logistic Regression	EOF FS	0.93	0.94
Decision Tree Classifier	EOF SMOTE	0.86	0.88
Random Forest Classifier	EOF FS	0.93	0.91
Gradient Boosting Classifier	EOF FS	0.94	0.92

Conclusions

Examining the relationship between EOF student retention and a specific predictor provided insights about common characteristics or patterns exhibited within the student population. Based on the results from Chapter 3, at ages 18 and 19 years old, EOF students are most susceptible to attrition. In particular, students who are majoring in psychology, biology, social work, communication, and marketing may not be retained. Looking at this more broadly, students who are in the Schools of Social Science and Human Services, Theoretical and Applied Science, and the Anisfield School of Business have the highest count of attrition. When considering gender, females have a higher count for attrition, but the males actually have a higher percentage rate of attrition than females.

Since this study defined retention as staying enrolled a year from their first semester, i.e., freshman year fall to sophomore year fall semester, or freshman year spring to sophomore year spring semester, it is obvious that freshmen students have the highest attrition rates. Similarly, students who are classified as new first-time students, which typically indicates a freshman student or a transfer student, have higher attrition rates. When examining residency status, EOF residents had higher retention and attrition rates than EOF commuters, which was surprising. Finally and unsurprisingly, students who retained had a higher average term GPA and average cumulative GPA than students who did not retain did.

Within chapter 4, where I created my own version of a report card for the EOF student population, students struggled the most in math 108, interdisciplinary study 101, biology 221, critical reading and writing 102, amer/intl interdisciplinary 201, math 101, and math 110. Generalizing these results, the EOF department may want to provide additional support and

personalized resources to students who are enrolled in math, biology, interdisciplinary studies, psychology, or chemistry courses.

Conducting a similar analysis exploring the relationship between retention/attrition and creating a report card for EOF students who majored in STEM, revealed that students majoring in biology and computer science have the highest attrition rates. Furthermore, the results from Chapter 5 suggest that the EOF should provide additional resources for students who are enrolled in math 108, biology 221, math 221, math 101, math 104, chemistry 116, and math 110. More generally, plans should be devised to support students who are enrolled in math, biology, chemistry, and computer science courses.

When implementing k-means clustering on all three datasets; all EOF students, all EOF students pre-covid, and all EOF students post-covid, the highest silhouette score was associated with two clusters. The clusters typically were not that distinct from each other when analyzing retention, on-campus residency status, and school. The primary difference between the clusters was that one cluster always had significantly more students than the other, for all EOF students and all EOF students post-covid this was labeled cluster 1.

Comparing the recommended models for predicting EOF first-year retention, the logistic regression, decision tree classifier, random forest classifier, and support vector machine methods concur that based on the metric Precision (NR), the best model is the *EOF FS & SMOTE*. However, the gradient boosting classifier and ensemble methods favored the *EOF SMOTE* as the recommended model. Even though all six methods did not agree on one specific model, all six of the recommended models did implement SMOTE, suggesting the importance of having a balanced dataset. Based on the results within Table 13.1, the recommended model is the random forest classifier *EOF FS & SMOTE*.

Expanding upon the results in Table 13.1, the results in Table 13.2 showed that for three out of the four recommended models that implement feature selection, the logistic regression, decision tree classifier, and random forest classifier methods all found the feature matric term significant.

In the future, if I had more time to work on this project, I would adapt the recommended random forest *EOF FS & SMOTE* model to explicitly tell them whether or not the student will retain along with the probability of retention. The EOF department could then use its expertise to develop a personalized plan or commit resources to support those students most at-risk of dropping out. The real test to see if the model was effective would be looking at the retention and EOF graduation rates a year later. Specifically, did they improve, stay the same, or decrease? Long-term, I would update this model for the departments continuously, as the more data the model has, the better it can be generalized in the future.

It would also be beneficial to expand the STEM analysis to track those students that are retained but switch out of a STEM major. There are possible STEM attrition issues that are not captured by our analysis but could use similar methods to understand the scope of the STEM pipeline problem at Ramapo College and predict persistence in those majors.

References

Alam, Mohammad Arif Ul. “College Student Retention Risk Analysis from Educational Database Using Multi-Task Multi-Modal Neural Fusion.” *arXiv.Org*, 11 Sept. 2021, arxiv.org/abs/2109.05178.

Alkhasawneh, Ruba; Hargraves, Rosalyn Hobson, *Journal of STEM Education: Innovations and Research*, v15 n3 p35-42 Oct-Dec 2014

Amelec Viloria, Jholman Garcia Padilla, Carlos Vargas-Mercado, Hugo Hernández-Palma, Nataly Orellano Llinas, Monica Arrozola David, Integration of Data Technology for Analyzing University Dropout,Procedia Computer Science, Volume 155, 2019, Pages 569-574, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.08.079

Aulck, Lovenoor, et al. *Mining University Registrar Records to Predict First-Year Undergraduate Attrition*, 2 July 2019, www.adobe.com/acrobat/pdf-viewer-extension.html.

Braunstein, Andrew W., Mary Lesser, and Donn R. Pescatrice. "The Business of Freshmen Student Retention: Financial, Institutional, and External Factors." *The Journal of Business and Economic Studies*, vol. 12, no. 2, 2006, pp. 33-53,75-76. *ProQuest*, http://library.ramapo.edu:2048/login?url=https://www.proquest.com/scholarly-journals/business-freshmen-student-retention-financial/docview/235795471/se-2.

Bringle, Robert G.; Hatcher, Julie A.; Muthiah, Richard N. *Michigan Journal of Community Service Learning*, v16 n2 p38-49 Spr 2010

“College Education Linked to Higher Pay, Job Security, Healthier Behaviors and More Civic Involvement: New College Board Report.” *College Education Linked to Higher Pay, Job Security, Healthier Behaviors and More Civic Involvement: New College Board Report*, College Board, 2 Jan. 2017, newsroom.collegeboard.org/college-education-linked-higher-pay-job-security-healthier-behaviors-and-more-civic-involvement-new.

Dursun Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems, Volume 49, Issue 4, 2010, Pages 498-506, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.06.003.

Delen, D., Davazdahemami, B. & Rasouli Dezfouli, E. Predicting and Mitigating Freshmen Student Attrition: A Local-Explainable Machine Learning Framework. *Inf Syst Front* (2023). https://doi.org/10.1007/s10796-023-10397-3

Fadel, Soufaine. “Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review.” *Explainable Machine Learning, Game Theory, and Shapley Values: A*

Technical Review, Government of Canada, Statistics Canada, 28 Feb. 2022, www.statcan.gc.ca/en/data-science/network/explainable-learning.

Fernandez, Alberto, et al. "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary." *View of Smote for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-Year Anniversary*, 20 Apr. 2018, jair.org/index.php/jair/article/view/11192/26406.

Géron, Aurélien. *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly, 2020.

"Home." *Educational Opportunity Fund Program*, 22 June 2023, www.ramapo.edu/eof-program/. *Data Science Process Alliance.* 28 April 2024. https://www.datascience-pm.com/crisp-dm-2/. Accessed 8 May 2024.

Hotz, Nick. "CRISP-DM". Photograph.

Matz, S.C., Bukow, C.S., Peters, H. *et al.* Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Sci Rep* **13**, 5705 (2023). https://doi.org/10.1038/s41598-023-32484-w

Olbrecht, Alexandre M.; Romano, Christopher; and Teigen, Jeremy (2016) "How Money Helps Keep Students in College: The Relationship between Family Finances, Merit-based Aid, and Retention in Higher Education,"Journal of Student Financial Aid: Vol. 46: Iss. 1, Article 2. Available at: http://publications.nasfaa.org/jsfa/vol46/iss1/2

Ramapo College of New Jersey: 2022 Fact Book, chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/www.ramapo.edu/ir/wp-content/uploads/sites/52/2023/09/2022-FACT-BOOK-_final_9.5.23.pdf. Accessed 19 Apr. 2024.

Tatiana A. Cardona, Elizabeth a. Cudney, Predicting Student Retention Using Support Vector Machines, Procedia Manufacturing, Volume 39, 2019, Pages 1827-1833, ISSN 2351-9789, https://doi.org/10.1016/j.promfg.2020.01.256.

Tinto, Vincent. "Student Retention and Graduation: Facing the Truth, Living with the Consequences. Occasional Paper 1." *Pell Institute for the Study of Opportunity in Higher Education*, Pell Institute for the Study of Opportunity in Higher Education. 1025 Vermont Avenue NW Suite 1020, Washington, DC 20005. Tel: 202-638-2887; Fax: 202-638-3808; e-mail: info@pellinstitute.org; Web site: http://www.pellinstitute.org, 30 June 2004, eric.ed.gov/?id=ED519709.

Trostel, Philip A. "It's Not Just the Money: The Benefits of College Education ..." *DigitalCommons@UMaine*, Lumina Foundation, 2015, www.luminafoundation.org/files/resources/its-not-just-the-money.pdf.

Videla, Nicole. *Exploring Collegiate Career Development Experiences of Educational Opportunity Fund (EOF) Alumni*, Northeastern University, United States -- Massachusetts, 2020. *ProQuest*, http://library.ramapo.edu:2048/login?url=https://www.proquest.com/dissertations-theses/exploring-collegiate-career-development/docview/2472095576/se-2.

Appendices