Predicting Determinants of Success Among College Students (Outline)

Samantha Gonzales, Jay Johnson, Alec Poss, Bibi Baarh, Laura Pineda, Aiden Gonzalez

STAT 520 Applied Multivariate Analysis

Professor Ryan Paul Lafler

October 29, 2025

Abstract

Our research aims to identify the socioeconomic, cultural, and institutional factors that most strongly influence levels of "success" among U.S. college students. Using longitudinal HERI data, we applied Alexander Astin's Input-Environment-Outcome (I-E-O) model to define the variables we would focus on (Astin & Antonio, 2012). We then constructed a condensed dataset composed of our 36 key variables within the Python framework. Normality and Ad Hoc tests are first conducted in SAS for our key variables to understand our metadata. Our analytic framework employs first a Confirmatory Factor Analysis (CFA), then a Structural Equation Modeling (SEM) analysis performed in R using the lavaan package. We believe this analysis, along with random forest and gradient boosting models, will allow us to find significant relationships among our key variables and constructs to determine associative and predictive strength. This work and our results found within aim to reveal the factors that are most associated with student success and, thus, can be utilized to aid in improving higher educational outcomes throughout the US.

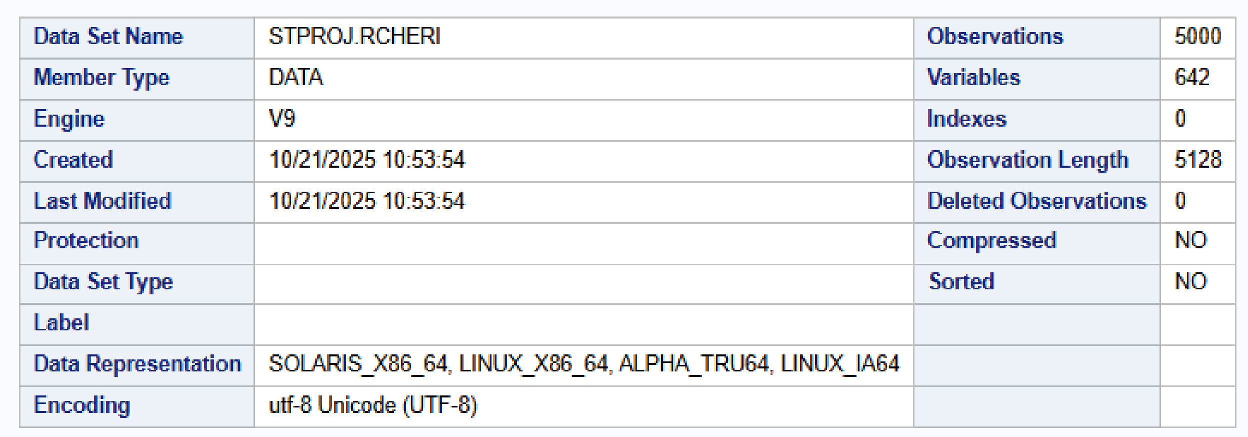
Introduction

According to the Education Data Initiative, "college enrollment totaled 19.28 million undergraduate students nationwide in Fall 2024" (Education Data Initiative, 2024). This number is not the highest enrollment, but undergrad enrollment has been increasing steadily for more than half a century. With this increased enrollment, it is important that all these students are able to find success in their college careers, both academically and professionally. Through our research, we aim to answer the question: What socioeconomic, cultural, and institutional factors best predict educational and career success among U.S. college students? The challenge we explore centers on college students who face barriers that hinder their success in achieving academic and long-term success, which we aim to examine and determine through our study. Our goal is to use spatiotemporal data to analyze which specific factors in a college student's education most influence their success. Our data source includes two datasets, longitudinal in design, from the 2000 College Freshman and the 2004 College Senior Survey instruments, administered by the Higher Education Research Institute (HERI) established at UCLA.

Add: methodologies, what programming language?

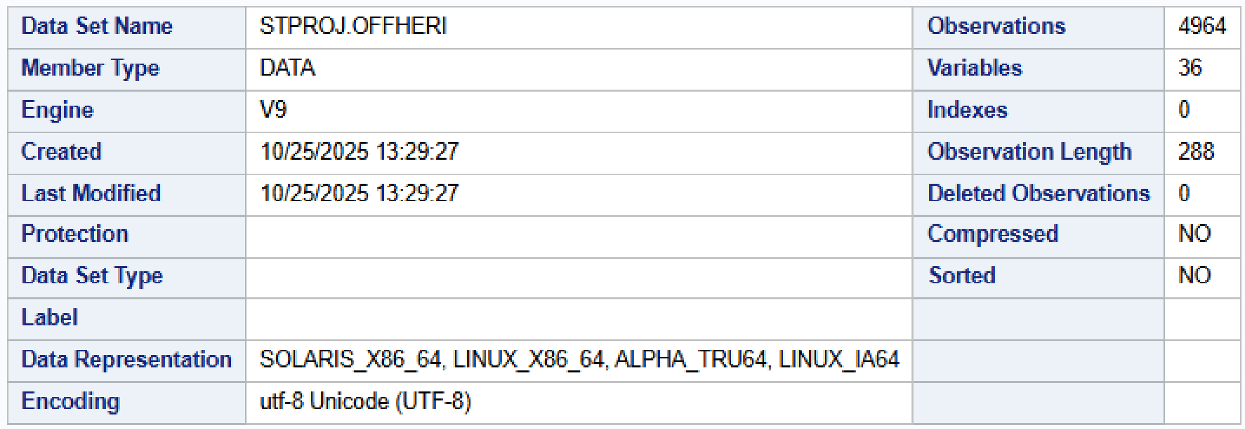
Data Sources Add: Data Characteristics

After a long search for a dataset that fit our goals in answering our research question, we identified two longitudinal datasets: the 2000 College Freshman Survey and the 2004 College Senior Survey from the Higher Education Research Institute (HERI). HERI is an educational research program founded in the 1960s, based at “the University of California, Los Angeles (UCLA), and, since 2023, has been jointly led by the UCLA School of Education and Information Studies (UCLA Ed&IS) and the American Council on Education (ACE)” (Higher Education Research Institute, 2024). The About section on the HERI’s website describes the Institute as “the longest running, most comprehensive data collection of institutes of higher education, including data on more than 1,900 institutions, over 15 million students, and more than 300,000 faculty” (Higher Education Research Institute, 2024). These surveys collect detailed information on student demographics, academic experiences, and personal development, capturing their attitudes and goals as they enter college in the Freshman Survey and measuring their individual academic progress, satisfaction, and outcomes in the Senior Survey. An important factor in our decision to choose this data is its longitudinal design: it records responses from the same students four years apart. By tracking the same students over time, the dataset helps control many external factors that might be present in a non-longitudinal study of similar interest.



Data Set Name	STPROJ.RCHERI	Observations	5000
Member Type	DATA	Variables	642
Engine	V9	Indexes	0
Created	10/21/2025 10:53:54	Observation Length	5128
Last Modified	10/21/2025 10:53:54	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Our raw dataset included both the 2000 Freshman Survey and the 2004 Senior Survey in a single large SAV file. After importing this dataset into SAS, the use of a PROC CONTENTS procedure produced a metadata table (shown above), which revealed 5,000 observations and 642 variables in which each observation was recorded. When confronted with this massive dataset, our first action was to find a way to narrow our focus to a smaller sample of the 642 variables. In doing this, we made our data easier to work with while still keeping our research question in mind. To achieve this, we employed the IEO model, created by Alexander Astin in his efforts to study the effects college has on students (Astin & Antonio, 2012). This model allowed us to go column by column and, with our own judgment, assign variables to four categories: Independent Variables (Input), Predictor Variables (Environment), Dependent Variables (Output), and Success Variables (Output).



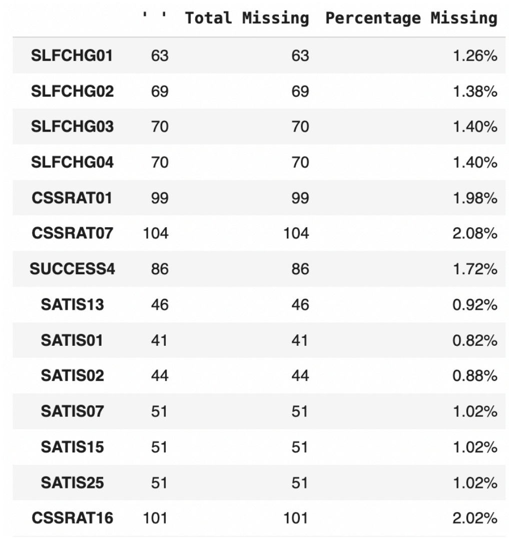
Data Set Name	STPROJ.OFFHERI	Observations	4964
Member Type	DATA	Variables	36
Engine	V9	Indexes	0
Created	10/25/2025 13:29:27	Observation Length	288
Last Modified	10/25/2025 13:29:27	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Now, with this shortened list, we were able to create a new dataset with just 36 variables (shown above) that align most closely with our unique focus while retaining nearly all 5,000 observations to maintain data reliability for our later predictive statistics and machine learning (ML) model.

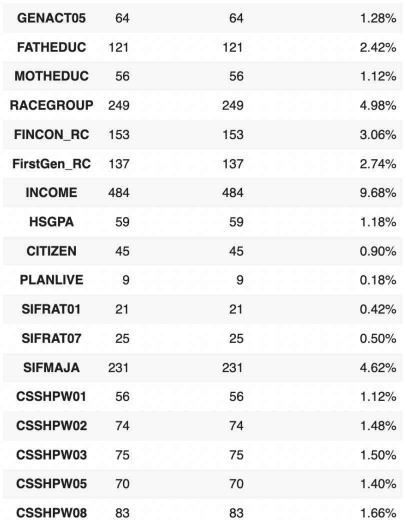
Methodology

Data Cleaning

To begin our data cleaning, as described above, by using the IEO model to help us in variable selection, we cut down from 642 variables to 36 relevant variables that ultimately bring the most significance to our question. To explore missingness, we first aimed to address the observations (rows) that were missing more than 40% of the data. We did not want to impute those rows' data as it could potentially bias our results by using participants that were made up heavily of imputed data. After adding this code, we removed 36 observations, resulting in a total of 4964 participants.

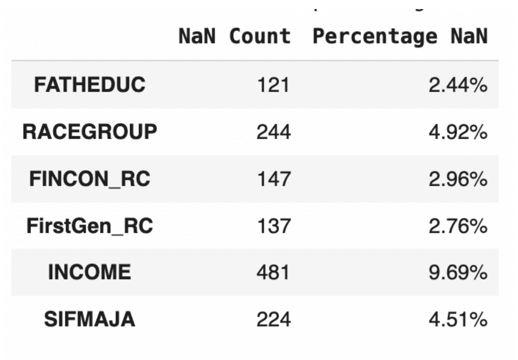


	' '	Total Missing	Percentage Missing
SLFCHG01	63	63	1.26%
SLFCHG02	69	69	1.38%
SLFCHG03	70	70	1.40%
SLFCHG04	70	70	1.40%
CSSRAT01	99	99	1.98%
CSSRAT07	104	104	2.08%
SUCCESS4	86	86	1.72%
SATIS13	46	46	0.92%
SATIS01	41	41	0.82%
SATIS02	44	44	0.88%
SATIS07	51	51	1.02%
SATIS15	51	51	1.02%
SATIS25	51	51	1.02%
CSSRAT16	101	101	2.02%



GENACT05	64	64	1.28%
FATHEDUC	121	121	2.42%
MOTHEDUC	56	56	1.12%
RACEGROUP	249	249	4.98%
FINCON_RC	153	153	3.06%
FirstGen_RC	137	137	2.74%
INCOME	484	484	9.68%
HSGPA	59	59	1.18%
CITIZEN	45	45	0.90%
PLANLIVE	9	9	0.18%
SIFRAT01	21	21	0.42%
SIFRAT07	25	25	0.50%
SIFMAJA	231	231	4.62%
CSSHPW01	56	56	1.12%
CSSHPW02	74	74	1.48%
CSSHPW03	75	75	1.50%
CSSHPW05	70	70	1.40%
CSSHPW08	83	83	1.66%

Next, we created a table that demonstrated the percentage of missingness per variable located above. For variables with under 2% of missing values, we decided to use a simple imputation method, either the mode or median, depending on the variable type. Mode is acceptable for nominal variables and median is acceptable for ordinal variables because nominal data have categories without order, while ordinal data have a ranked order that allows for identifying a middle value. Below are the variables that were not imputed using these methods.



	NaN Count	Percentage NaN
FATHEDUC	121	2.44%
RACEGROUP	244	4.92%
FINCON_RC	147	2.96%
FirstGen_RC	137	2.76%
INCOME	481	9.69%
SIFMAJA	224	4.51%

These variables were missing more than 2% of data, so the code did not impute these variables. To fix this missingness, we first ran a Little MCAR (Missing Completely at Random) Test for the 6 variables above. This test checks if the data is missing at random, missing not at random, or missing completely at random. Our output stated that these variables are not missing at random, concluding that we need a more complex imputation method. We decided on the Multiple Imputation by Chained Equations (MICE) method. The MICE method imputes by using the values that are present in each row to predict the missing ones through iterative regression modeling.

This is an approved imputation method for MAR, which we are assuming because of the small percentages of missing. This resulted in zero missing or empty values, and with this, we saved and created a new dataset to further explore for the rest of our research.

Exploratory Data Analysis and Statistical Modeling

In the exploratory phase, we first plan to inform the construction of latent variables, which represent broader psychosocial and institutional constructs that cannot be measured directly (Kline, 2023). Indicators for these constructs (i.e., satisfaction, sense of belonging, faculty interaction) will be drawn from the HERI 2000-2004 surveys (HERI, 2004). Exploratory factor analysis (EFA) will first be conducted to identify the underlying structure of the survey items and determine factor loadings above 0.40 (Fabrigar & Wegener, 2012). Items that show cross-loadings or weak communalities will be removed before confirmatory factor analysis (CFA) is conducted to validate the factor structure.

Next, CFA will be used to establish construct validity, reliability, and measurement invariance across demographic subgroups (Byrne, 2016). Each latent construct *(socioeconomic capital, institutional engagement, cultural influence, and perceived success*) will be validated using goodness-of-fit indices such as CFI ≥ .95, TLI ≥ .95, RMSEA ≤ .06, and SRMR ≤ .08 (Hu & Bentler, 1999). Once the measurement model is confirmed, the structural equation model (SEM) will estimate the directional relationships among both observed indicators and our latent constructs. SEM will allow for the simultaneous modeling of direct, indirect (mediated), and moderated effects (Hayes, 2022).

Analysis

Following model validation, the baseline structural model will be specified with *socioeconomic capital* predicting *academic success* through *Institutional Engagement* and *cultural influence* as mediators. This approach will allow for testing both direct and indirect pathways of influence (Kline, 2023). The SEM will be estimated in R using the lavaan package with maximum likelihood estimation and bootstrapped standard errors, preferably at 95%. Associate models (i.e., partial mediation, full mediation) will be compared using χ^2 difference tests and information criteria (AIC, BIC).

Machine learning methods will complement the SEM analysis by assessing predictive strength rather than causal direction. Random forest and gradient boosting algorithms will be used to estimate variable importance and nonlinear interactions among predictors (Breiman, 2001). Dimensionality reduction through PCA will precede clustering analysis to uncover distinct student profiles, which may reflect the latent constructs identified in the SEM. Evaluation metrics for supervised models will include R^2, RMSE, and MAE, while unsupervised models will be assessed using silhouette scores and within-cluster variance (James et al., 2021).

Initial Hypothesis (Hypothesized Results)

The team hypothesizes that socioeconomic status and cultural capital will indirectly influence academic success through institutional engagement and academic self-efficacy. Specifically, students with higher socioeconomic resources and stronger engagement indicators will report greater academic confidence and higher GPA by senior year. The hypothesized mediation model will be validated through SEM, and its predictive parallels will be tested through machine learning algorithms to assess the convergence between explanatory and predictive approaches.

Discussion and Industry Applications

Our topic focuses on identifying the factors — demographic, economic, emotional, etc. — that lead to greater success in a student's continued education. Studies like ours have wide-ranging applications—from informing government initiatives aimed at improving national educational attainment and competitiveness among OECD countries, to aiding individual schools in fostering an environment that helps their students get the most out of their time there. Beyond our analysis, our work can also help parents and educators recognize the shared factors that the most successful college students experience — mainly social and emotional ones —to foster a home and school environment that can set students up for greater success.

Conclusion and Future Work

Our work in this project highlights the strengths of longitudinal data for studying and analyzing patterns within the social sciences. The ability to observe changes in the same people over time allows us to have greater confidence in our results and findings. Newer HERI data could be used to continue our work and be implemented in our model to find trends among generations of college students. To expand the scope of this research, HERI and similar programs could conduct studies of graduates' early-career outcomes after college, allowing for meaningful comparisons of success across institutions and degree programs.

Further, our research could be used around college campuses to aid the creation of inclusion programs and inform policies. Based on our findings, if specific groups are seen as more or less successful than others, programs can be implemented that are based on this research in order to target communities that need more assistance and make our country's education system as a whole more equitable.

Authors' Bios

Bibianca Baarh

Fourth-year Information Systems major with a minor in Statistics. Part-time Strategy intern at Lansweeper Inc., where I support the Mergers and Acquisitions team by conducting market research on emerging tech trends and collaborating with the product team to identify potential targets for acquisitions.

Alec Poss - alecposs1212@gmail.com

Senior undergraduate student at San Diego State University, majoring in Economics and minoring in Statistics. Alongside my Economic research, my work in Statistics has enabled me to develop skills in programming languages such as SAS. I am currently on track to graduate in December 2025 and aim to find a position where I can apply my skills in data analysis alongside my expertise in Economics, both nationally and internationally.

Aiden Gonzalez - aideng324@icloud.com

Senior undergraduate student at San Diego State University, majoring in Statistics with a minor in Mathematics. As a part of my major and minor I have also gained experience in Economics and basic coding skills in Java, Python, R, C++, and SAS. I am currently on track to graduate in December 2025 and looking into entry level positions with both private banking companies and local government.

Laura Pineda - laurap5750@gmail.com, LinkedIn

Fourth-year undergraduate student studying Applied Mathematics with a minor in Statistics at San Diego State University. Working as a Math and Statistics tutor at the Math & Science Learning Center at SDSU, focusing on all levels of math. Experienced in various programming languages, including R, Python, SAS, and Java. Graduating

Spring 2026, with the hopes of starting a Master's program focusing on Applied Mathematics.

Samantha Gonzales - sgonzales2392@sdsu.edu, LinkedIn
Senior undergraduate student at San Diego State University, majoring in Sociology and minoring in Statistics. My coursework has focused on research design, data analysis, and social inequality. I have previously used Python for school projects. I also interned at the National Immigration Law Center, where I led data projects to assess campaign engagement. I also led the Dignified Learning Project on campus, where I oversaw event planning for a student research conference.

Judd (Jay) Johnson - jjohnson4039@sdsu.edu,

Jay Johnson is a doctoral candidate in the Administration, Rehabilitation, and Postsecondary Education program at SDSU. His dissertation examines how underrepresented and first-generation students with dual enrollment credits adjust emotionally, socially, and academically during their first year. He is a graduate of UCLA (B.A.) and the University of Florida (M.F.A.) in Performing Arts, English, and Communications. Beyond academia, he is also an accomplished television and film actor with more than 25 national commercial credits, including campaigns for Apple, Toyota, Verizon, and Southwest Airlines. His professional interests merge data analytics, educational equity, and creative pedagogy. He has presented at national conferences, including sessions on applied neurology and movement-based pedagogy, which introduced arts based learning strategies to STEM faculty. Jay's technical experience includes statistical modeling in SPSS and Mplus, along with growing proficiency in Python for data handling and visualization. His long term goals include becoming a Vice President for Academic Affairs while continuing to publish as an education researcher.

References

· American Council on Education. (n.d.). *American Council on Education.* https://www.acenet.edu/Pages/default.aspx

· Astin, A. W., & Antonio, A. L. (2012). “A conceptual model for assessment” (Chapter 2). In *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education.* Rowman & Littlefield.

https://www.vumc.org/faculty/sites/default/files/Assessment/AstinAlexanderW_2012_Chapter2_AssessmentForExcellen.pdf

· Astin, A. W., & Oseguera, L. (2012). Pre-college and institutional influences on degree attainment. In A. W. Astin (Ed.), *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (3rd ed., pp. 239–263). Rowman & Littlefield.

· Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

· Byrne, B. M. (2016). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (3rd ed.). Routledge.

· Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.

· EducationData.org. (2025). *College enrollment statistics.* Education Data Initiative. https://educationdata.org/college-enrollment-statistics

· Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis.* Oxford University Press.

· Field, A. P. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage.

· Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

· HERI. (2004). *Higher Education Research Institute: 2000 Freshman and 2004 Senior Surveys* [Data set]. University of California, Los Angeles.

· Higher Education Research Institute. (2024). *About HERI.* University of California, Los Angeles. https://heri.ucla.edu/about-heri/

· Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

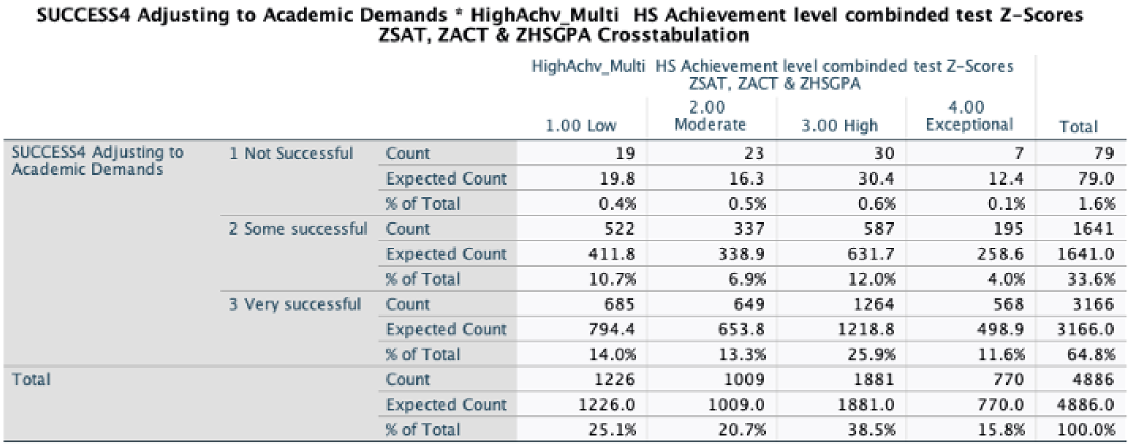
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

- Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

- UCLA School of Education & Information Studies. (n.d.). *Homepage.* https://seis.ucla.edu/

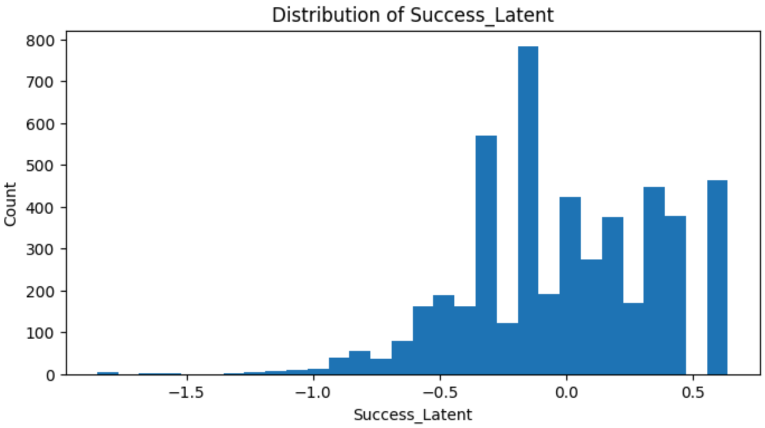


SUCCESS4 Adjusting to Academic Demands * HighAchv_Multi HS Achievement level combinded test Z-Scores ZSAT, ZACT & ZHSGPA Crosstabulation

			HighAchv_Multi HS Achievement level combinded test Z-Scores ZSAT, ZACT & ZHSGPA				
			1.00 Low	2.00 Moderate	3.00 High	4.00 Exceptional	Total
SUCCESS4 Adjusting to Academic Demands	1 Not Successful	Count	19	23	30	7	79
		Expected Count	19.8	16.3	30.4	12.4	79.0
		% of Total	0.4%	0.5%	0.6%	0.1%	1.6%
	2 Some successful	Count	522	337	587	195	1641
		Expected Count	411.8	338.9	631.7	258.6	1641.0
		% of Total	10.7%	6.9%	12.0%	4.0%	33.6%
	3 Very successful	Count	685	649	1264	568	3166
		Expected Count	794.4	653.8	1218.8	498.9	3166.0
		% of Total	14.0%	13.3%	25.9%	11.6%	64.8%
Total		Count	1226	1009	1881	770	4886
		Expected Count	1226.0	1009.0	1881.0	770.0	4886.0
		% of Total	25.1%	20.7%	38.5%	15.8%	100.0%

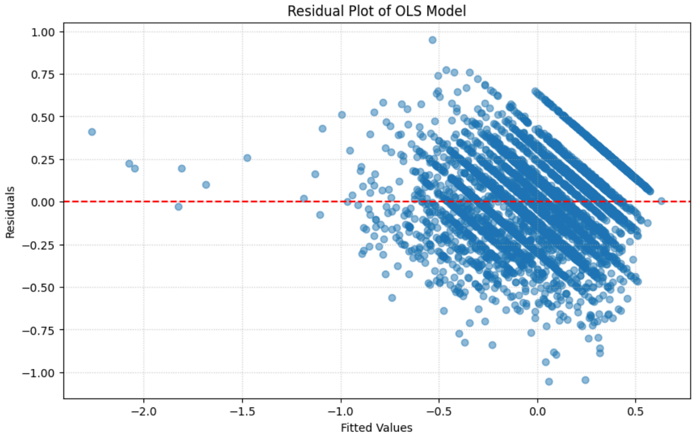
Machine Learning Models

To incorporate Machine Learning into this project, conceptually, we wanted to be able to predict the new success latent variable. The priority was to predict Success in order to capture which predictors had the most significance and impact on the response. To begin, we made a histogram, shown below, of the success variable in order to visualize the outcomes' spread, shape, and any outliers. The histogram revealed no outliers that were drastically impacting the shape of the distribution. The shape appeared roughly symmetrical and bell-shaped, however, there was a slight left skew. Overall, the Success variable followed an approximately normal distribution.



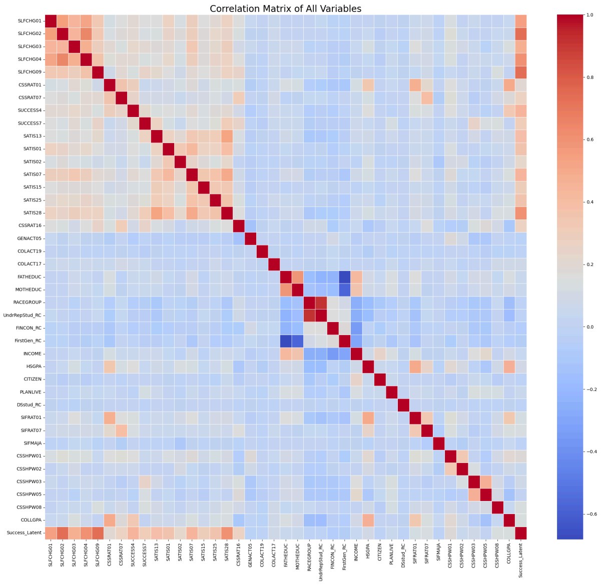
OLS - Ordinary Least Squares

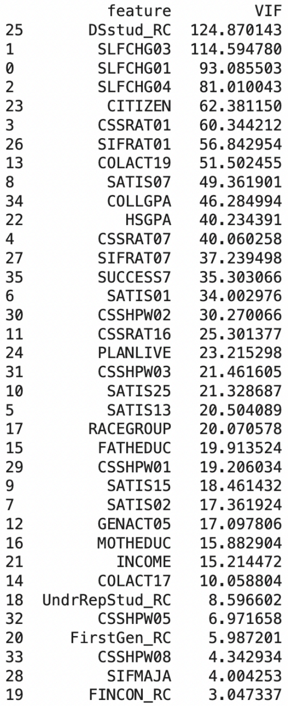
Next, to start incorporating machine learning models, we began with OLS Regression. For all models created, the predictor set did not include the variables used to create the success latent variable. This is the foundational linear model, and we started with it in order to create a baseline model and understand the relationship further between the predictors and the response. The OLS model had an R-squared of .592 and an adjusted R-squared of .589. For social science data, this is generally a good fit. The F-statistic was very high, 198.8, and Prob (F-statistic) was extremely low, essentially 0. This means that at least one of the predictors is significantly related to our success variable. However, looking at the p-values of each of the predictors, they varied between significant and not significant. Looking further into the tests performed on the model, the Omnibus and Jarque-Bera tests both came out with very low values, indicating the residuals are not normally distributed, which will affect the accuracy of p-values. Then, the Condition Number came out to be 652, which suggests extremely high multicollinearity among predictors. Both of these findings were worrisome, so we decided to explore these issues further. Below is a residual plot created to interpret if they were distributed normally and confirm what was found in the model summary.



The residual plot indicates that the residuals are not evenly distributed around zero. There is a visible issue and possible heteroscedasticity, suggesting that the assumptions of the OLS model may not be fully satisfied.

Further, we wanted to explore the correlation between all of the variables and the multicollinearity between predictors. Below is a correlation matrix representing the relationships among predictors and a VIF chart presenting those values that represent collinearity. Ideally, these values should be below 10 for no impact on the model.





```
           feature         VIF
25       DSstud_RC  124.870143
1         SLFCHG03  114.594780
0         SLFCHG01   93.085503
2         SLFCHG04   81.010043
23         CITIZEN   62.381150
3         CSSRAT01   60.344212
26        SIFRAT01   56.842954
13        COLACT19   51.502455
8          SATIS07   49.361901
34         COLLGPA   46.284994
22           HSGPA   40.234391
4         CSSRAT07   40.060258
27        SIFRAT07   37.239498
35        SUCCESS7   35.303066
6          SATIS01   34.002976
30        CSSHPW02   30.270066
11        CSSRAT16   25.301377
24        PLANLIVE   23.215298
31        CSSHPW03   21.461605
10         SATIS25   21.328687
5          SATIS13   20.504089
17       RACEGROUP   20.070578
15        FATHEDUC   19.913524
29        CSSHPW01   19.206034
9          SATIS15   18.461432
7          SATIS02   17.361924
12        GENACT05   17.097806
16        MOTHEDUC   15.882904
21          INCOME   15.214472
14        COLACT17   10.058804
18  UndrRepStud_RC    8.596602
32        CSSHPW05    6.971658
20     FirstGen_RC    5.987201
33        CSSHPW08    4.342934
28         SIFMAJA    4.004253
19       FINCON_RC    3.047337
```

The map and table above demonstrate the strong multicollinearity among multiple predictors. In the correlation matrix, you can see clusters of variables that are highly correlated. These include DSstud_RC, SLFCHG indicators, and CSSRAT, meaning they can be measuring similar things and are constantly moving in the same direction. High correlation can sometimes indicate multicollinearity. Therefore, after creating the VIF table, you can see that the conclusion transfers over. Consequently, these variables are causing an unstable pattern in our coefficients for our model. This can cause inflated standard errors, weak statistical accuracy, and non-interpretable findings. Our goal of the project is to uncover which predictors have a high impact on success, therefore, this model is not reliable. Due to these problems, we chose to move forward with Random Forest, Lasso, Ridge, and Elastic Net models instead. These models and methods are more robust to multicollinearity. Random Forest reduces correlated predictor influence through ensemble averaging, while Lasso, Ridge, and Elastic Net apply regularization to shrink or eliminate redundant coefficients, improving model stability, interpretability, and predictive performance.

For each of the models, the dataset was randomly split into an 80/20 training and testing set. This was to be sure our models were the least biased as possible. The dataset is very large, as previously stated, so this split has enough information for both the testing and training sets. Additionally, all models used a 5-fold cross-validation method in order to make sure that we capture the average performance of each model, rather than overfitting and making our conclusions unreliable. All of these methods were chosen and the same for all models to make sure comparison was fair.

Random Forest

First, we incorporated Random Forest because it was a non-linear aspect we wanted to add, as it performs well even when there is multicollinearity. The random forest model's metrics were a test MSE of 0.0582, a test R-Squared of 0.5792, and an OOB R-Squared of 0.5551. These are strong averages as social science data is hard to capture with close to complete accuracy. The most influential predictors were SLFCHG04, SLFCHG01, SATIS13, SATIS07, SIFMAJA, and INCOME. Interestingly, Random Forest was able to capture the importance of income and major (SIFMAJA), unlike the other models, suggesting they may not have a linear relationship with Success, but are still significant.

Lasso Regression

Next, we included a Lasso Regression model, which performs variable selection through the penalty of variables with no significance. The model chose an alpha of 0.00195 and had the metrics of a test MSE of 0.0559 and an R-Squared of 0.596. The variables Lasso thought were most significant were SLFCHG04, SLFCHG01, SATIS13, SATIS07, SUCCESS7, SLFCHG03, and CSSRAT16. These variables suggest that Lasso found that students' self-reported change and satisfaction were huge contributors to their success.

Ridge Regression

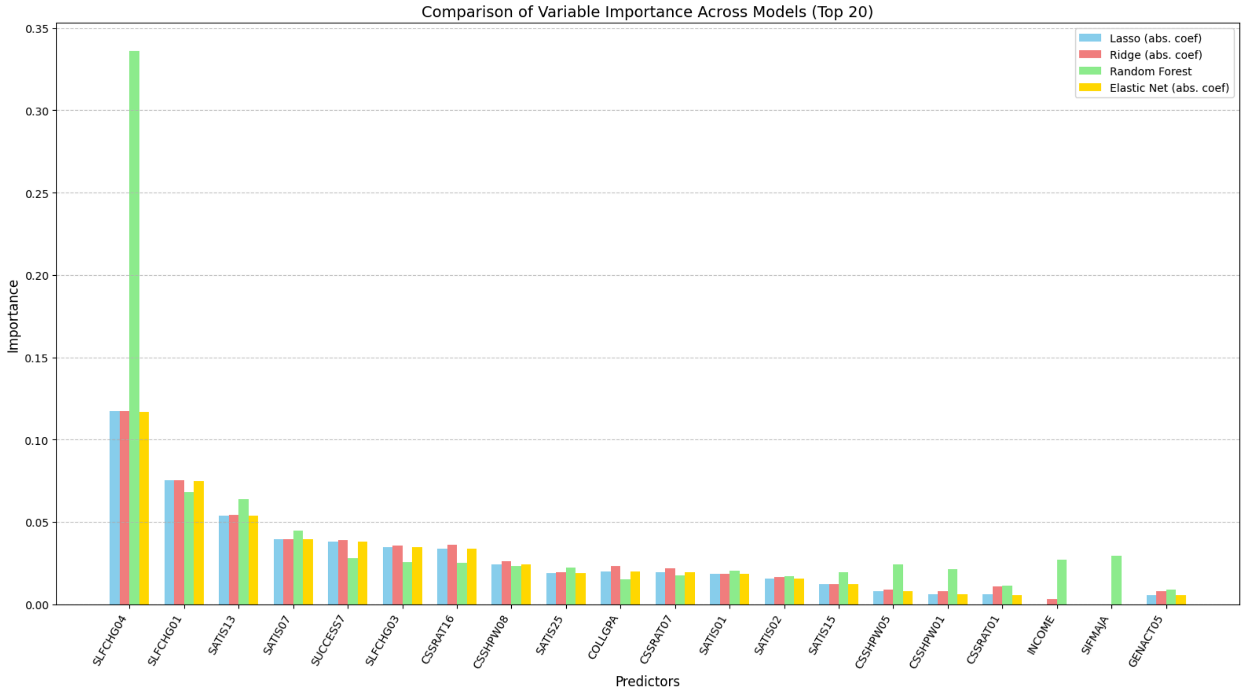
Then, Ridge Regression is a little different than Lasso and more lenient because it does not automatically shrink coefficients to zero and perform variable selection, it just shrinks coefficients, but still keeps them all in the model. Ridge is good for multicollinearity because it stabilizes the coefficients, but does not remove the variables entirely. Ridge selected an alpha of 10 after cross validation. The R-Squared came out to 0.59595, which was almost the exact same as Lasso. This model found that SLFCHG04, SLFCHG01, SATIS13, SATIS07, and CSSRAT16 were the most significant predictors, all chosen by Lasso as well, indicating that these are constantly being portrayed as important variables.

Elastic Net Regression

Lastly, Elastic Net Regression was used as it combines Lasso and Ridge by using both forms of penalty. The chosen alpha was 0.00390 and the test R-Squared was 0.59596, showing consistency between all regression models. This model showed the same group of variables Lasso and Ridge did as most significant.

Model Comparisons and Important Predictors

Overall, the goal of our project was to assess which variables had the most impact on academic performance. In our case, we measure success through our latent variable of success and then evaluate the impact that predictors have on it through the outcomes of different machine learning models. The bar chart below demonstrates the overlap each of the models had and which variables they found significant. The most influential predictors were SLFCHG04 (Critical Thinking Growth), SLFCHG01 (General Knowledge Growth), and SLFCHG03 (Field Knowledge Growth). This revealed that, according to the machine learning models, self perceived growth and development was strongest in predicting their success.



The self reported Satisfaction variables also ranked high amongst models. Specifically, SATIS13 (Sense of Community), SATIS07 (Quality of Instruction), SATIS25 (Ability to Find Faculty or Staff). This shows that community and environment are highly significant as well.

Additionally, Random Forest found nonlinear significance in Income and SIFMAJA (Major).