College Scorecard Graduation Rate Analysis: ML Final Paper

By: Selah, Siobhan, Eva, Lilly, Noyanika, Janna, and Rachel

Abstract

The rising cost of college tuition and increasing student debt have made the value of a college degree a pertinent topic. To address this, our team used graduation rates as a proxy for the value of attending specific institutions. We applied both supervised and unsupervised machine learning methods to determine the most significant predictors of graduation rates. Using data from the U.S. Department of Education, we employed ordinary least squares (OLS) regression for supervised learning and principal component analysis (PCA) for secondary, unsupervised analysis. Our findings revealed three key variables strongly associated with graduation rates: the percentage of students taking loans, average ACT scores, and institutional region. The percentage of loans serves as a proxy for financial burdens on students, average ACT scores reflect students' academic preparedness, and institutional regions highlight broader socioeconomic factors such as regional income and education levels. Additionally, our analysis showed that OLS regression outperformed PCA. While PCA reduced dimensionality and offered more insight than multicollinear data, OLS offered slightly superior predictive accuracy and greater interpretability, making it the preferred method for this study. In conclusion, financial, academic, and regional factors have an influence on graduation rates, and OLS regression proves to be an effective tool for identifying these relationships.

Introduction

In this study, we explore the factors that influence graduation rates at U.S. higher education institutions, using graduation rates as a proxy for the "value" of attending those institutions. This approach stems from an understanding that academic success is not always solely the product of individual effort or intelligence, but rather the result of complex interactions between various financial, academic, institutional, and demographic factors. With increasing emphasis from employers on the importance of a college degree for career entry or advancement coupled with the rising costs of tuition nationwide, it is crucial to analyze what truly drives student success. This helps ensure that higher education provides a meaningful return on investment & is structured with an adequate support network to allow for equitable educational and career outcomes for all students, regardless of their background.

The U.S. Department of Education's College Scorecard 2022-2023 dataset serves as our primary data source. This dataset contains detailed information on institutional characteristics, financial aid, and student outcomes, including variables such as tuition costs, student loan debt, family income, enrollment size, and, most importantly, graduation rates.

Our aim is to identify the key predictors of graduation rates and understand how these factors interact with one another to influence student success. Specifically, we focus on three main areas:

1. Financial Factors: How the cost of attending college and the financial aid available to students (e.g., the percentage of students receiving loans or Pell Grants) affect their ability to graduate.

2. Academic Factors: The relationship between students' academic preparedness, as measured by standardized test scores like the ACT, and their graduation success.
3. Institutional and Demographic Factors: The influence of institutional characteristics, such as the type of school (public vs. private), and regional disparities in education quality and access.

Prior to more advanced analysis, we created various visualizations plotting the relationship between different variables within the dataset. Prominent graphs included a scatterplot depicting cost of attendance vs graduation rate, boxplots depicting graduation rate by region and average earnings by predominant degree, a histogram displaying the distribution of ACT midpoint scores, and a scatterplot depicting Pell Grant recipients vs. graduation rate organized by school type. These visualizations gave us a foundational understanding of the dataset; they were a means for us to communicate our initial findings effectively, situate later analysis within the broader context of our study, and ultimately help us to identify variables of interest that warranted deeper investigation. They also highlighted other variables that, while less influential, could still contribute to graduation rate outcomes, potentially indicating areas for future research/exploration.

We also applied both supervised and unsupervised machine learning methods to explore these factors. For supervised learning, we used Ordinary Least Squares (OLS) Regression, a statistical method that allows us to examine the relationship between predictor variables (percent of students receiving loans, average ACT scores) and graduation rates. This method not only provides direct, straightforward insight into the strength of these relationships but also allows us to quantify the contribution of each variable to the overall prediction of graduation rates. For unsupervised learning we used Principal Component Analysis to look at our multivariable

dataset as a small set to analyze clusters or trends of the top components that influence graduation rates(average loans, average ACT, and school region). The combination of both of these learning methods allowed us to gain a more comprehensive understanding of the factors influencing graduation rates; OLS provided clear, easily interpretable insights into the impact of specific predictors on graduation rates, while PCA helped us identify overarching trends & analyze the broader structure of the dataset while reducing dimensionality of the data.

Data

The College Scorecard 2022-2023 dataset provides detailed information on higher education institutions across the United States. When we downloaded the data from the US Department of Education, the file was too large to upload to Github so we chose specific variables of interest and combined them into a dataset to use for this project. It includes data on institutional information, student aid, enrollment, cost, and student outcomes. By offering insights into these factors, the dataset helps assess the accessibility, affordability, and effectiveness of universities. This dataset is a valuable resource for analyzing patterns in higher education, allowing us to investigate factors that influence student success and performance.

Initially, we planned to use the College Scorecard 2022-2023 dataset to investigate whether college could be considered a "scam" and identify factors contributing to the perceived value of higher education. However, after testing the data and receiving feedback, we revised our approach to focus on creating a predictive model for the U.S. Department of Education. The primary objective of the model is to identify institutions and measures that fail to support student success, allowing the Department of Education to implement targeted interventions and improve graduation rates.

We carefully reviewed the dataset, retaining only the variables most relevant to our project. Variables were renamed for clarity to better reflect what they represent and ensure consistency in interpretation. One challenge we encountered was understanding the variables and what they measure. The data dictionary, while comprehensive, could be confusing at times, but a careful review allowed us to select the most relevant variables for our goals. We referenced the data dictionary provided by the Department of Education to understand the sources of missing data. According to the dictionary, many NA values were due to institutions not reporting certain metrics, either for privacy reasons or lack of available information. To ensure the quality of the dataset, we removed rows with missing values in critical variables, creating a clean and usable dataset for analysis.

By analyzing graduation rates (GRAD_RATE, renamed to graduation_rate), the model can determine which types of institutions—based on affordability, demographics, and financial aid—are more likely to struggle with supporting students to graduate. Variables like percent loans (PCTFLOAN_DCS_POOLED_SUPP, renamed to percent_loans) and Pell Grant recipients (PCTPELL_DCS_POOLED_SUPP, renamed to pell_grant) can help evaluate how well colleges serve underrepresented or economically disadvantaged populations. Comparing median earnings after graduation (MD_EARN_WNE_4YR, renamed to average_income) to tuition costs (COSTT4_A, renamed to yearly_cost) enables an assessment of whether graduates earn enough to justify the cost of their education. School type (CONTROL_PEPS, renamed to school_type) allows for comparisons between public, private nonprofit, and for-profit institutions, shedding light on how institutional type influences affordability, student debt, and graduation rates. Geographic data such as region (REGION) reveals regional disparities in access to affordable education and varying graduation outcomes. Ultimately, by examining how colleges operate

(e.g., prioritizing profit over education), the analysis raises questions about whether current systems perpetuate inequities and prioritize neoliberal values, such as commercialization and privatization, over educational quality.

Methods

The team aimed to answer the main question of "how do socioeconomic factors, particularly average family income, predict graduation/completion rates at US colleges?" The plan was to focus on C150_4 (completion rate for first-time, full-time students at four-year institutions within 1.5x expected completion) as the single target variable to narrow our analysis and allow for more discernment to which variables are most strongly associated with higher C150_4 values. The team hypothesized that lower average family income, higher average cost of attendance, higher DCS federal loan rate pooled, and higher DCS Pell grant rate pooled will be correlated with lower completion rates. This question is significant considering the tension between American ideals of "pulling yourself up by your bootstraps" and hidden structural barriers that influence educational outcomes & social mobility; exploring how differing financial backgrounds might lead to educational disparities can prompt deeper reflection on the notion that success is solely based on individual effort and illustrate the need for equitable policy that enables student success across socioeconomic divisions.

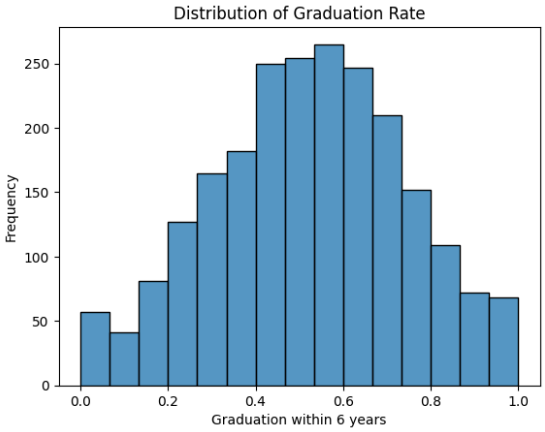
All of our analyses and visualizations were performed in Google Colab (Python) with the Numpy, Pandas, MatPlotLib, and Seaborn packages loaded to perform the analysis and create the visualizations.

The team created two different models utilizing a combination of supervised and unsupervised learning. The supervised learning component consisted of OLS regression to predict student outcomes. Then we used LASSO in order to ascertain what factors most

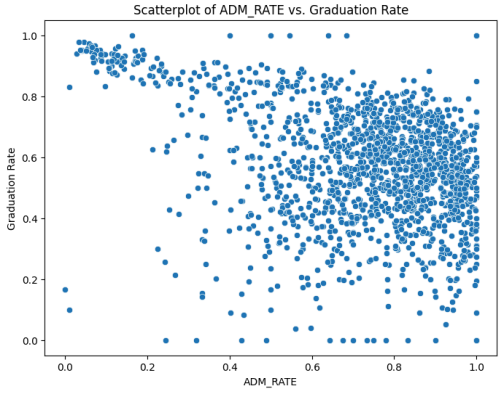
accurately predict successful student outcomes such as graduation, and to build a model based on the most predictive variables.

The second model utilized Principal Component Analysis to create proxies for the groups of variables in buckets (those representing socioeconomic status as well as school prestige) and incorporated those into a linear regression model.

Based on the histogram to the right we can see that our target variable, graduation rate, is continuous, numeric, and the distribution seems relatively unimodal without skew, which indicates it could work for linear regression. Because of this, and because the vast majority of our variables are numeric we believe regression would be the most suitable tool for our analysis.



The team utilized multiple linear regression using the ordinary least squares (OLS) method to quantify the predictor values on C150_4 (completion rate for first-time, full-time students at four-year institutions). This will allowed the team to determine which coefficients to use for each variable (i.e. Average Family Income Percentage Student receiving Pell Grant etc).



This tool was chosen because it is easily interpreted and it will allow us to explore interactions between the variables as well as incorporating nonlinear, higher power terms into the model. For instance the relationship between admission rate and graduation rate (shown in the

scatter plot below) appears to be non-linear, so the team attempted to incorporate this into the model.

To build this model, a 80/20 train test split was created, and from that the model was built using the training data. This test data was then put aside to serve as a control to prevent and identify data overfitting.

In addition to the supervised learning method above where we include all variables that seem relevant, the team realized that various explanatory variables initially set to build the model fell into two main categories: measures of average socioeconomic status at the school (Average Family Income, Percentage of Students receiving Pell Grants etc.) and measures of prestige of the school (Midpoint ACT score, Admission Rate etc.). Because these variables are largely attempting to measure similar things, using all of them in the model would likely lead to multicollinearity and overfitting. The team pivoted to make another model using the unsupervised learning algorithm of Principal Component Analysis to create proxies for the groups of variables in buckets (those representing socioeconomic status as well as school prestige). Then the resulting buckets were compared to the model trying to use them all, comparing both against the test set. The problem with this approach however is that this process reduces explainability of the model because it is difficult to make sense of what the transformed variables really mean. This method is also susceptible to outliers and overfitting.

To know if our approach works we used the data from the train test split to test our model while reducing the risk of overfitting the model to the data it was trained on. Since the data was split into training and testing data, the team first trained both models with the training data and then assessed the approach using the testing data set. More specifically, the R squared and RMSE

of the test set with the trained model. The team then determined which model is best based on which performs better with the test set.

Linear regression of graduation rate on multiple variables such as acceptance rate and income was then run and R squared and RMSE was calculated to see how accurate these variables are in predicting the graduation rate. The resulting linear regression of graduation rate on acceptance rate, act score, and cost of attendance, the R squared score was 0.739 and RMSE: 0.0957. So this shows that these variables have a high rate of prediction and it has a low error in prediction. Just with this the model seems successful, because it is more accurate than not, however more tests were run to confirm these results; therefore, these results were compared head to head with PCA analysis.

Residual analysis was also performed with the difference between observed and predicted calculated and plotted against the predicted values.This will allow us to check for patterns related to linearity or potential outliers; if the residuals are all randomly scattered around zero, the model indicates a linear relationship.

The weakness of the OLS regression model is the high number of variables that could potentially come into play and isolating the individual or power of one variable is also difficult due to the fact that these variables are also highly correlated. For this reason, the team used PCA to deal with multicollinearity in such a high dimensional data set.

In the case the approach fails, it indicates the presence of additional variables not accounted for (omitted variables bias) or a focus on the wrong variables . These numerical variables like act score, cost of attendance, acceptance rate, have been accounted for , but the one categorical variable was also taken into account and proved to be a good predictor, public or private school, that skewed the RMSE and R squared.

To prepare the data for our analysis we first pre-processed by handling the missing data, identifying and managing the outliers, and transforming these categorical variables. Additionally, the variable on the most common degree awarded (PREDDEG) narrows the data frame to schools that primarily offer undergraduate degrees to increase similarity and comparability across our observations. One hot coding is used for binary categorical pieces of data. For example the variable on whether type of school; public, private nonprofit, or private for-profit, (CONTROL_PEPS) would be good candidates for one-hot encoding since they are binary, categorical pieces of data. Region would also need to be one hot encoded. Secondly, for the many correlated numerical variables, PCA was performed to reduce multicollinearity. An example of such variables would be PCTFLOAN_DCS_POOLED_SUPP and PCPPELL_DCS_POOLED_SUPP in order to provide insight into financial assistance and socioeconomic demographics between students and how this would relate to variables such as college completion rate.

The results were communicated with head to head comparisons of test-set r-squared and RMSE from each of the models to communicate accuracy of the models. Regression coefficients were presented via a table to enhance explainability and understandability of the model, such that those viewing the results can understand how each explanatory variable plays into the model.

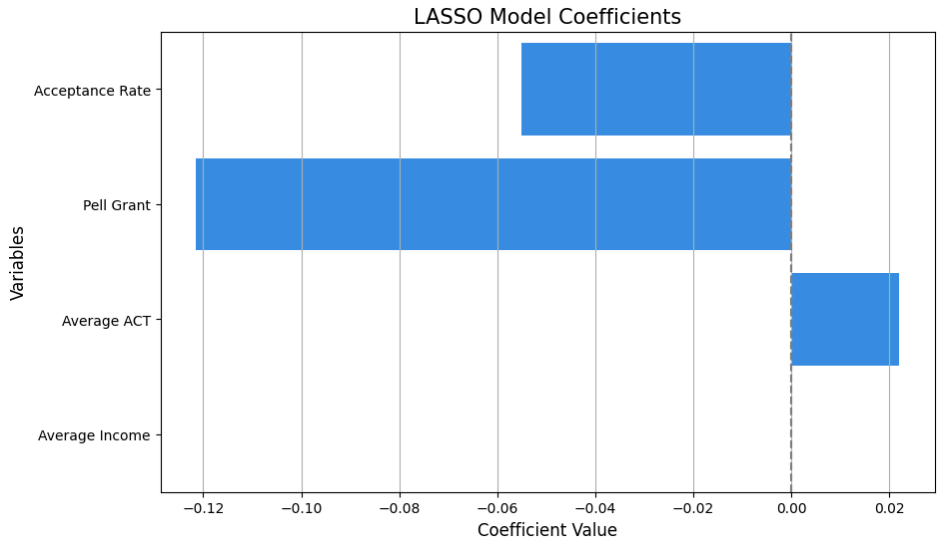
These models compare the correlation between socioeconomic variables like average family income, average cost of attendance, DCS federal loan rate pooled, DCS Pell grant rate pooled, and completion/graduation rate in a table. Higher correlation values would indicate stronger associations between the predictor variable and completion rate, indicating higher predictive power for the variable. Visualizations based on the correlation matrix, such as a

heatmap, for more effective visual summarizing/communication and to make pattern observation easier.

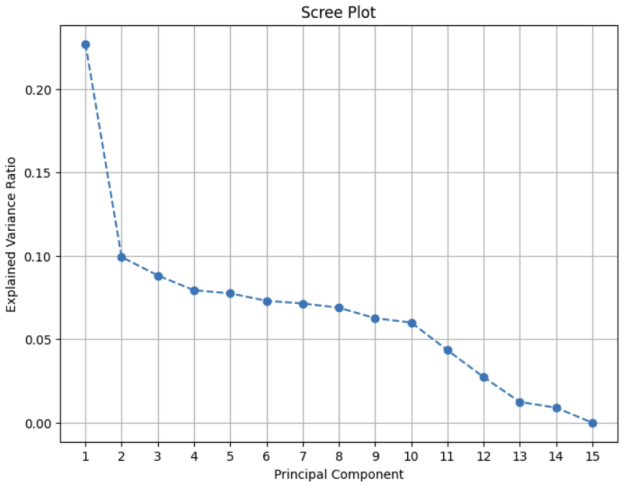
Results

Our prediction question was: "What socioeconomic and institutional factors are most influential on college graduation rates in the US?" We wanted to build a model to help the Department of Education identify key factors that could guide targeted interventions to improve student outcomes. By identifying these influential factors, the DOE could allocate resources more effectively to the institutions that require the most support.

Our group worked with two models, the first being an OLS model. OLS (Ordinary Least Squares) regression is a statistical method used to estimate the relationship between a dependent variable and one or more independent variables. In our analysis, we used OLS to examine the relationship between various parameters and their influence, or lack thereof, on graduation rates.



This graph shows the coefficients from the lasso regression model. The bars represent the magnitude and direction of the coefficients for each variable which indicate how changes in each variable affect graduation rates when other factors are held constant. In this model, as the acceptance rate increases, graduation rates decrease. This could suggest that schools with less selective admissions policies may have lower completion rates, potentially due to students not being prepared. The model also returned a negative correlation for the Pell Grant variable, possibly reflecting the challenges lower-income students face. As shown in the graph, acceptance rate and Pell Grant are the most influential factors. This could help to guide the DOE in creating targeted interventions. Initiatives such as increasing support for Pell Grant recipients and increasing academic support at less selective institutions could be implemented based on the model.



Top Contributing Features for Each Principal Component:

	Feature	Loading
0	average_act	0.492848
1	percent_loans	0.488067
2	region_southeast	0.474217
3	region_mid-east	0.703403
4	region_plains	0.765086
5	region_far-west	0.646704
6	region_far-west	0.594367
7	region_US-service-schools	0.730210
8	region_US-service-schools	0.654540
9	stud_fac_ratio	0.664941
10	average_income	0.419602
11	acceptance_rate	0.599918
12	average_act	0.851469
13	pell_grant	0.705316
14	region_southeast	0.503835

After applying PCA, we analyzed the loadings for each principal component to find the features that contributed the most to the observed variance in the data. The scree plot revealed a clear elbow at two components, indicating that the first two capture the majority of the variance in the data. Further analysis found that 11 principal components explain about 95% of the variance in the dataset; we came to the conclusion that while the two component model provides

a good summary, including more components allows for a more comprehensive, holistic understanding of the data. The table above highlights the top contributing features for each principal component. Average ACT score and percent loans contribute most heavily to the first two components, while components 2-8 are primarily driven by regional indicators, likely reflecting the geographical disparities in education outcomes. Additional components include contributions from features like student-faculty ratio, acceptance rate, and Pell Grant percentages, which may reflect the importance of institutional support, inclusivity, and accessibility for students educational outcomes.

From this analysis, we determined that the percent of loans, average ACT, and school region best predict graduation rate among the studied factors. These features likely are broader indicators of student characteristics: percent loans likely show financial burdens faced by a student, average ACT score perhaps shows at a surface level the academic preparation of incoming students, and institutional region shows the geographical setting/living situation of students.

Model	R-squared	RMSE
OLS (Lasso using all variables)	0.7657	0.0817
PCA	0.7552	0.00698

When comparing the two models used, the OLS model was the best model. For this model, the R squared value, an indicator of the percentage of the variance that can be explained by the model is about **76.5%** which indicates that the model can explain a significant portion of the variability. The low RMSE value at 0.0817 indicates a low average difference between predicted and actual values, indicating a reasonably low error for the model. The linear regression model identifies specific variables that are correlated with graduation rates and

provides coefficients that are straightforward to interpret as opposed to PCA. Lasso served as a means of dealing with the inherent limitations and pitfalls of linear regression by providing feature selection.

The PCA model was slightly less effective, with an R squared value of 0.7552–explaining about **75.52%** of the variance in graduation rates. We also thought its lower RMSE value compared to the OLS model might mainly reflect differences in the scale of output rather than greater accuracy. This model excels at reducing dimensionality, making it useful when dealing with multicollinearity among predictors and summarizing the data. However, it lacked the straightforward interpretability of the OLS model; principal components were combinations of variables and the loadings must be analyzed to identify/understand the contribution of individual features to variance.

Conclusion

In this project, we set out to answer the question "What socioeconomic and institutional factors are most influential on college graduation rates in the US?" This is significant because understanding these factors can guide data driven policies that support student success and improve national education outcomes. Using OLS and PCA models, we studied the relationship between various factors, such as average family income, region, and graduation rates. In particular, we used these models to quantify the impact of certain variables on completion rates so that we could provide a comprehensive resource for schools and institutions that want to raise graduation rates. Our OLS model, which was slightly more accurate and explained 76.5% of the variation in data, determined that a low acceptance rate, low percentage of Pell Grant students, and high ACT scores are the most influential factors for a higher graduation rate. The PCA

model (which explained 75.52% of the variation), highlighted ACT scores and federal loan rates as the most critical factors, with regional differences also playing a big role.

While we considered the factors with the greatest weight in each of the derived principal components to address our research question, our results are limited by the presence of multicollinearity. For example, factors such as average family income and percentage of Pell Grant recipients overlap, making it hard to determine their individual effects. Further, factors like median ACT score are also affected by other variables since students from wealthier backgrounds have access to tutoring services or better K-12 school systems to prepare them for standardized tests and college success. In short, using PCA to reduce dimensionality makes analysis and interpretation more straightforward, but it limits the conclusions that can be drawn for individual factors. By recombining our large set of variables into a smaller set of uncorrelated components, we lose some of the ability to draw conclusions about particular factors and their individual impact on graduation rate.

The dataset itself may have limitations that also influenced our findings. For example, this dataset integrates data from multiple sources, primarily the IPEDS system, which relies on self-reported data provided by institutions themselves. This may result in inconsistencies across the data, with smaller schools potentially lacking the capacity to report their data as comprehensively or accurately as larger schools with more resources. The data also represents a single academic year (22-23), which is a limited snapshot of institutional performance that lacks the context of broader trends/changes over time. It also focuses on broader institutional-level data rather than individual characteristics like mental health or motivation, which constrains analysis of nuanced interactions between these factors that impact academic outcomes.

Beyond the limitations of the models themselves, our results should not be used in isolation but instead should serve as a starting point for deciding how to increase completion rates. For instance, the OLS model exposed a negative correlation between completion rate and the percentage of students who are on Pell Grant. If someone was only looking at this result, they would conclude that the federal government should provide fewer Pell Grants in order to increase national graduation rates. However, any reasonable person would understand that this conclusion is invalid; since Pell Grant recipients are low-income, there are several confounding factors that decrease graduation rates. Instead of limiting the Federal Pell Grant program, the DOE should focus on reducing barriers for low income students, since it is clear that income is a key factor in determining likelihood of graduation.

To combat these limitations, this study could be conducted with more nuanced data that quantifies a student's resources both before and during college. For example, instead of just considering a university student's family income, quantifying access to test preparation, hours spent working during the school year, or availability of family support could offer deeper insights into the socioeconomic factors affecting graduation rates. Additionally, we could explore the impact of university-provided resources like free tutoring, counseling and mentorship opportunities, or mental health resources. An interesting extension could be to solely focus on these factors that universities have the power to control. Namely, instead of focusing on student background which may be difficult to control, by isolating the effects of institutional interventions, policymakers and administrators could identify effective strategies with the potential for immediate impact on graduation rates.

In conclusion, this project found key predictors of graduation rates, including acceptance rates, Pell Grant percentages, and average ACT scores. Our findings provide a foundation for

understanding how institutional and socioeconomic factors influence student success. While multicollinearity and data limitations presented challenges, these constraints offer opportunities for future research. By exploring factors and interventions that universities can directly influence, such as mentorship programs or tutoring, future analyses can build on our results to design targeted interventions that promote equity and improve graduation outcomes nationwide.

References

The U.S. Department of Education. (2024). *Data Home*. U.S. Department of Education College Scorecard. https://collegescorecard.ed.gov/