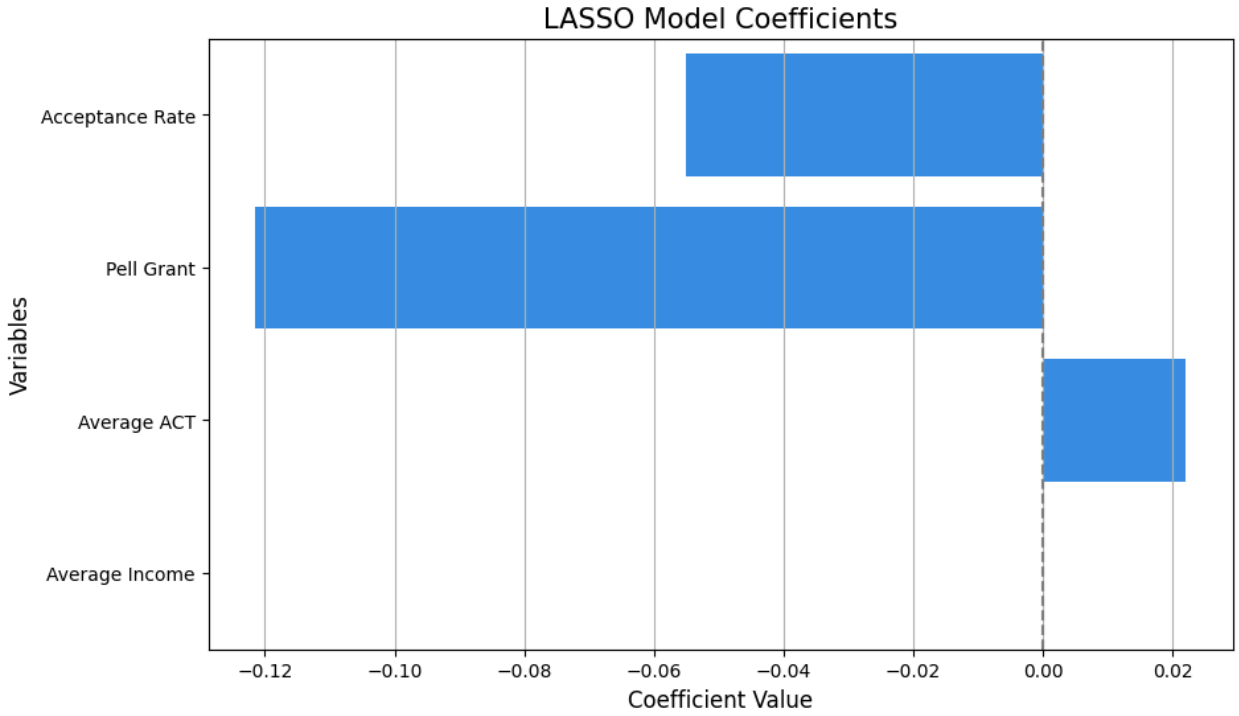
**Results**

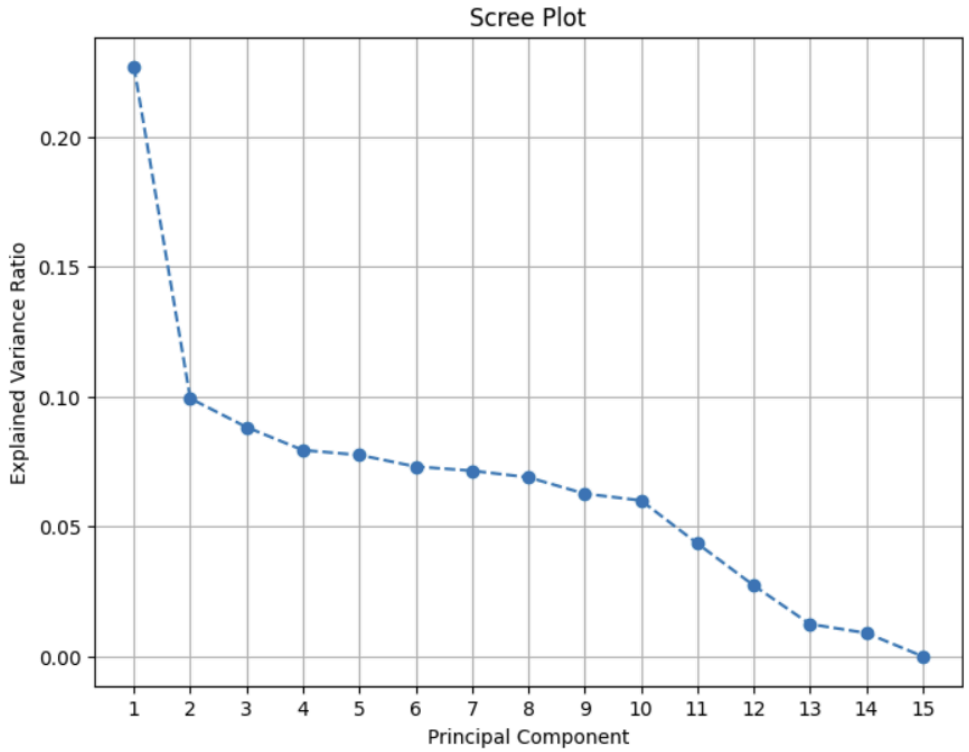
Our prediction question was: “What socioeconomic and institutional factors are most influential on college graduation rates in the US?” We wanted to build a model to help the Department of Education identify key factors that could guide targeted interventions to improve student outcomes. By identifying 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 |

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.