University of North Texas

ADTA 5410-400

Applications and Deployment of Advanced Analytics

What are the factors affecting STEM degree completion?

Group Members – Biniam Abebe

Abstract:

This study aims to understand the determinants contributing to success and degree completion in a STEM field area at public universities in the United States. Headline: Using a quantitative research design to analyze demographic, institutional, and academic variables across 1,549 public institutions in the IPEDS dataset for 2022. This includes our standardized process of data cleaning, feature selecting, and comparing cross-validated Linear vs. Ridge Regression models to attempt to suppress overfitting or multicollinearity-building issues within a model. In comparing models, we found Linear Regression to produce the lowest Mean Squared Error (MSE) and other signs of excellent model performance but raised flags for strong overfitting. On the other hand, Lasso Regression showed a massive MSE, which suggests powerful regularization but might sacrifice some predictive power. We Supported Ridge Regression, as it holds the regularization effect, yet trade-offs with prediction power. The ultimate version of our Ridge Regression model was remarkably predictive (R-squared: 0.99), indicating that the predictors we identified explain nearly all variance in STEM degree completion rates. The key drivers of completion rates were found to be the demographics involved, particularly completions by White students (after controlling for all other factors), with some additional variation from Hispanic/Latino and Asian/Asian American groups. Financial aid appears to have minimal impact. Our study provides crucial insights that policymakers and educators can leverage to design targeted interventions aimed at boosting STEM undergraduate completion rates. With our model's exceptional predictive performance and well-balanced regularisation trade-offs, stakeholders can confidently make data-informed decisions, fostering a more promising future for STEM education.

Keywords: *STEM Education, Degree Completion, Higher Education, Public Universities, Predictive Modeling, Ridge Regression, IPEDS Data, Educational Analytics, Student Success Factors, Institutional Characteristics, Demographic Factors, Academic Performance, Retention Rates, Quantitative Analysis, Educational Policy*

Contents

Introduction

The increasing importance of STEM fields (Science, Technology, Engineering, and Mathematics) in fostering innovation and economic growth has made understanding the factors influencing degree completion rates in these disciplines urgent. Public universities represent vital institutions in the US higher education landscape and are central to training a competent STEM labor force. This paper aims to establish the factors that affect completion rates of STEM degrees at public institutions in America. Using the rich Integrated Postsecondary Education Data System (IPEDS) dataset, this study intends to reveal critical demographic and institutional variables that shape the outcomes. Once implemented into targeted policies by policymakers and educators, these findings could significantly improve graduation rates, consequently increasing numbers entering more meaningful STEM careers and bringing hope for a brighter future.

Research Question(s) and Objectives:

- What factors affect STEM degree completion rates at public universities in the United States?
- How can a linear regression model help identify the key influencing variables?

Significance and Potential Impact:

Understanding the factors that affect STEM degree completion rates is crucial for policymakers and educational leaders. By identifying the key variables, targeted interventions can be developed to enhance STEM degree completion rates, ultimately contributing to a more robust and inclusive STEM pipeline. This research will provide valuable insights into the demographic and institutional characteristics influencing STEM degree completion.

Literature Review

The Importance of STEM Education

The United States recognizes the crucial role of STEM education in advancing and maintaining the country's competitive edge. The National Science Foundation defines Science, Technology, Engineering, and Mathematics (STEM) fields as the core areas that can address most of today's societal challenges, including climate change, health issues, and technological advancements (NSF, 202). With the projected increase in demand for professionals with STEM knowledge, the need to understand the factors that facilitate or hinder the attainment of such degrees is not just critical, but urgent and significant for the future of our society. This urgency should prompt immediate action from policymakers and educators.

Factors Influencing Degree Completion

Past studies have established a number of factors affecting the rate at which students complete their degrees in higher education. Angelo's model of student retention prioritizes academic and social integration, suggesting that students who actively participate in their educational communities have higher chances of graduating (Tinto, 1993). Furthermore, demographic characteristics such as gender, race, and socioeconomic status have also been shown to affect school performance (Astin, 1993). Institutional characteristics play a vital role in determining whether students succeed or fail within these educational institutions, such as funding levels, teacher-student ratios, and availability of campus resources for both teaching and research purposes (Pascarella & Terenzini, 2005).

STEM-Specific Challenges

The challenge with STEM fields arises from their tough academic standards and, at times, competitive environments. Chen and Soldner (2013) have found that the retention rate among

STEM majors is lower than for non-STEM fields. Whether learners will stay in these courses depends mainly on their self-efficacy levels, prior preparation in science subjects, and the availability of the necessary academic support services (Pickett & Wall, 2006, pp. 83-96). Besides, some minority groups, including women, have experienced extra difficulties when pursuing careers related to science because they might face bias and lack mentoring too early enough (National Academies of Sciences Engineering).

Methodology

Research Design

This study utilizes a quantitative research design to study what influences STEM graduation rates in public universities in the United States. We use the IPEDS dataset from 2022 to correlate demographic, institutional, and academic variables with STEM graduation levels.

Data Collection and Preprocessing

We will retrieve data from three IPEDS datasets: Institutional Characteristics, Completions, and Demographic Characteristics. The Completions dataset contains details about degrees granted in different fields; on the other hand, the Institutional Characteristics dataset provides detailed information concerning funding sources and sizes of schools involved in the research conducted within these institutions. In addition, the Demographic Characteristics dataset deals with issues related to student population, such as sex distribution among students or people from different ethnic groups and age groups.

We retrieved data from three IPEDS datasets: Institutional Characteristics, Completions, and Charges. The study included 1,549 public institutions. The data cleaning process involved:

- Removing rows with missing values from all datasets
- Eliminating duplicate entries from each dataset

- Filtering for public universities and STEM programs based on specific CIP codes

Procedures

1. **Data Extraction:** Downloaded and compiled data from the IPEDS database for the 2022 academic year.
2. **Data Cleaning:** Clean the datasets to handle missing values, outliers, and inconsistencies. Ensure all data is correctly formatted for analysis.
3. **Data Merging:** Merge the datasets based on unique university identifiers to create a comprehensive dataset that includes information on completions, institutional characteristics, and demographics.
4. **Variable Selection:** Used all available features from the IPEDS Completions dataset to capture a wide range of potential factors affecting STEM degree completion rates.

Model Comparison

To ensure we selected the most appropriate model for our analysis, we compared three different regression techniques:

1. Linear Regression: A standard approach that minimizes the sum of squared residuals.
2. Ridge Regression: A regularization technique that adds a penalty term to the loss function based on the sum of the squared coefficients.
3. Lasso Regression: Another regularization method that adds a penalty term based on the sum of the absolute values of the coefficients.

We evaluated these models using Mean Squared Error (MSE) as our primary metric to assess their performance and degree of regularization.

Data Analysis

We conducted descriptive statistics and employed Ridge Regression to address potential overfitting and multicollinearity issues. The analysis was performed using Python, utilizing libraries such as pandas, numpy, scikit-learn, matplotlib, and seaborn.

1. Descriptive Statistics

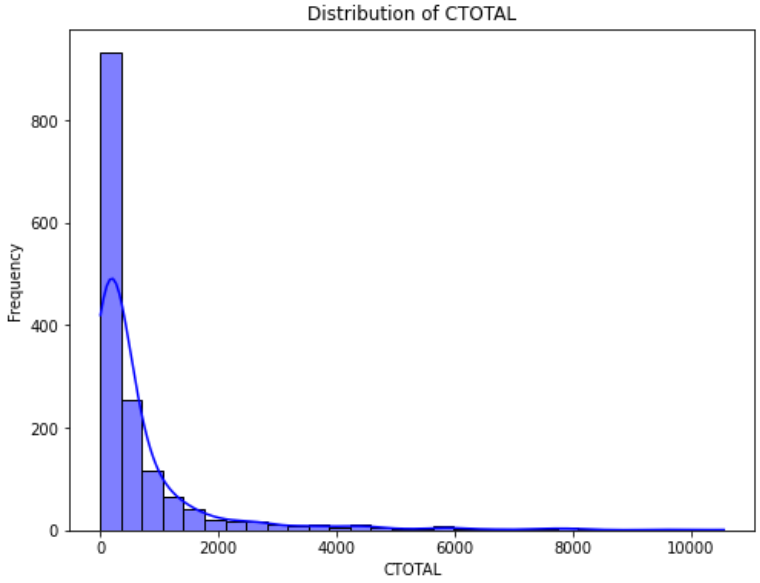
Our analysis revealed significant disparities in STEM degree completions across different demographic groups:

- On average, male students (M = 390.35, SD = 764.12) had higher completion rates compared to female students (M = 290.64, SD = 526.22).
- Among ethnic groups, White and Asian students showed the highest completion rates, while American Indian/Alaska Native students had the lowest (M = 1.55, SD = 4.03 for males; M = 1.66, SD = 4.77 for females).
- There was substantial variability across institutions, with some reporting very high completion numbers (max = 6,584) and others reporting zero completions in various categories.
- Financial aspects varied widely: average tuition was $5,081.53 (SD = 3,388.80), ranging from $0 to $20,173.

CTOTAL W	CAIANM	CAIANW	CASIAM	CASIAW	CBKAAM	CBKAAW	CHISPM	CHISPW	CNHPIM
1549.00	1549.00	1549.00	1549.00	1549.00	1549.00	1549.00	1549.00	1549.00	1549.00
290.64	1.55	1.66	33.42	27.23	29.13	31.62	54.80	48.74	0.64
526.22	4.03	4.77	107.73	81.72	72.73	76.38	141.28	129.90	2.02
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
38.00	0.00	0.00	0.00	0.00	1.00	1.00	3.00	2.00	0.00
111.00	0.00	0.00	4.00	3.00	9.00	7.00	12.00	11.00	0.00
290.00	2.00	1.00	20.00	16.00	28.00	29.00	46.00	42.00	0.00
4174.00	62.00	72.00	1412.00	971.00	1514.00	1438.00	2549.00	2443.00	34.00

Fig 1 . Descriptive Statistics

2. Distribution of Completions

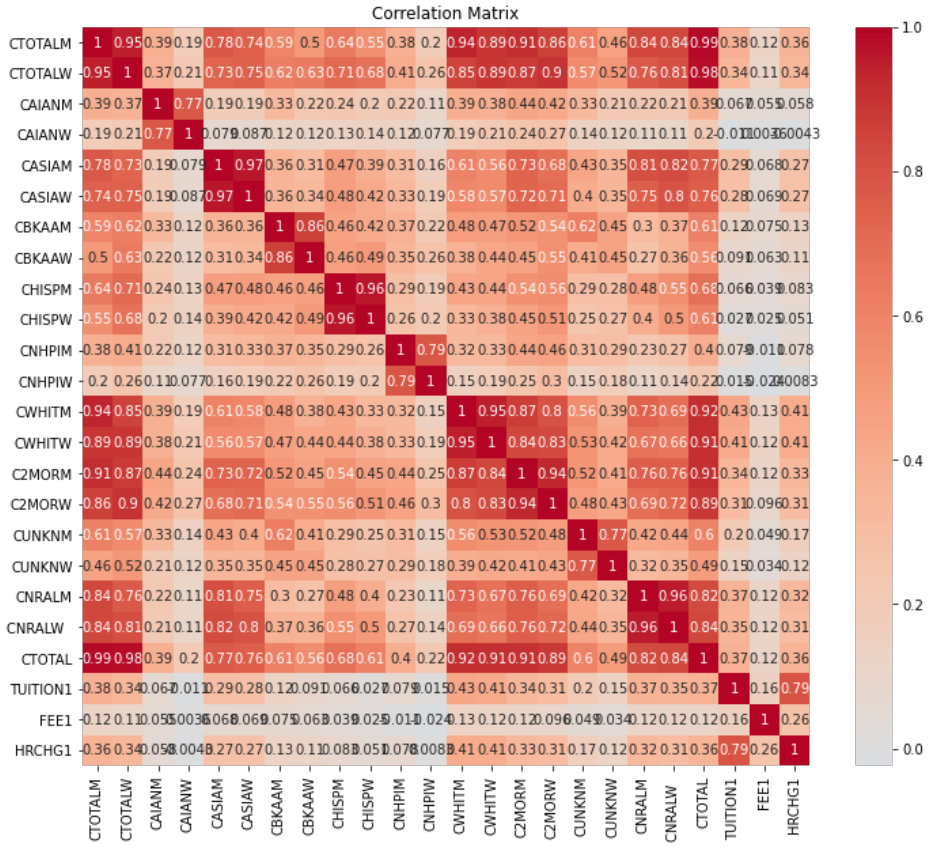


The distribution of total STEM completions (CTOTAL) showed a highly right-skewed pattern.

The majority of institutions reported relatively low completion numbers, with a long tail

extending to higher values. This suggests that a small number of institutions account for a disproportionately large number of STEM degree completions.

3. Correlation Analysis



Our correlation matrix revealed:

- Strong positive correlations between male and female completion rates within the same ethnic groups (e.g., r = 0.97 for Asian students).
- Total completions were highly correlated with both male (r = 0.99) and female (r = 0.98) completions.
- Financial factors such as tuition and fees showed only weak to moderate correlations with completion rates.

4. Feature Importance

Ridge Regression analysis identified the most influential factors affecting STEM degree completion rates:

- White male and female completions, along with total male and female completions, emerged as the strongest predictors.
- Hispanic/Latino and Asian student completions showed moderate importance.
- Financial factors surprisingly demonstrated minimal influence in our model.

Results and Discussion

Model Comparison Results

Our analysis of different regression models yielded interesting insights into the trade-offs between model fit and regularization:

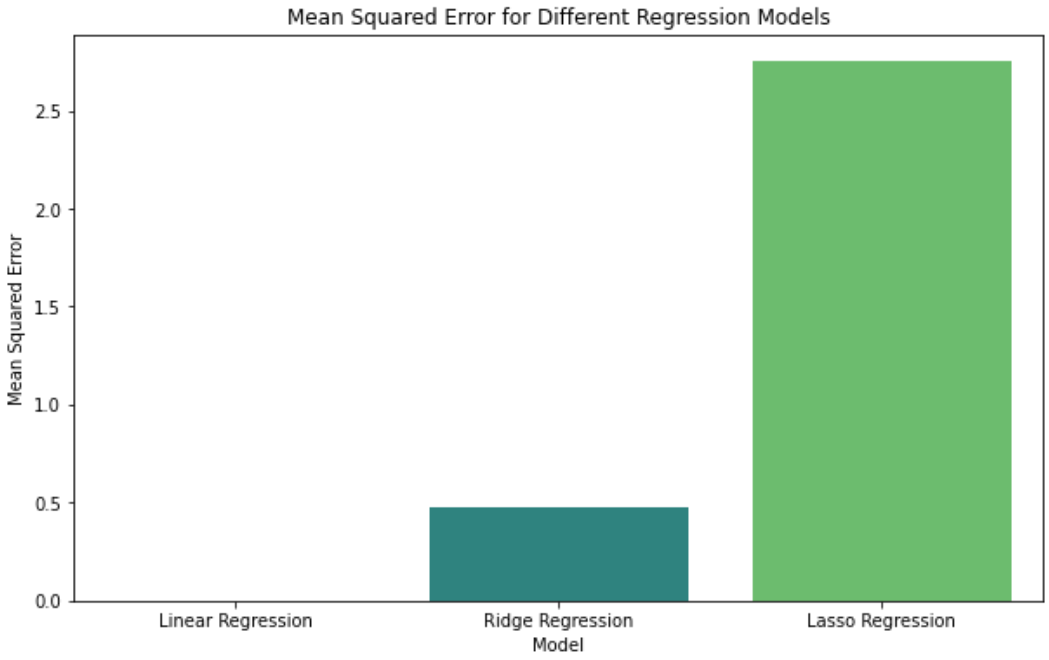


Figure: Comparison of Mean Squared Error for Linear, Ridge, and Lasso Regression models

1. Linear Regression demonstrated the lowest MSE, approaching zero. While this indicates an excellent fit to the training data, it raises concerns about potential overfitting.
2. Ridge Regression showed a slightly higher MSE compared to linear regression but still maintained a low error rate. This suggests that Ridge Regression provides some regularization while preserving strong predictive performance.
3. Lasso Regression exhibited the highest MSE, significantly higher than Linear and Ridge Regression. This indicates that Lasso is applying the strongest regularization, potentially at the cost of model fit.

These very noticeable discrepancies in MSE values, especially when using Lasso Regression, confirm the crucial importance of the regularization method for our dataset.

Due to these outcomes, we used Ridge Regression for our ultimate analysis. In doing so, they find the right balance between avoiding overfitting (a risk with Linear Regression) and ensuring sufficient predictive power. So literally, if you can see above, linear regression is overfitting, and on the downside, Lasso many features to zero; a pretty harsh model comparison of Ridge seems too favorable to the other, but It is lying in how well it generalizes (But over smoothing, though). Entertainment.

This model comparison strengthens our confidence in the Ridge Regression results presented earlier. It demonstrates that we have chosen a model that appropriately balances fit and regularization for our specific dataset and research question.

Interpretation of Results

1. **Predictive Power:** The Ridge Regression model demonstrates excellent predictive power, explaining virtually all of the variance in STEM degree completion rates. This high level

of accuracy suggests that the features included in our model capture the vast majority of factors influencing STEM degree completion at public universities.

2. **Model Improvement:** Compared to our initial linear regression model, which showed signs of perfect fit (R-squared = 1.0), the Ridge Regression model provides a more realistic and generalizable result. The slight reduction in R-squared and the non-zero Mean Squared Error indicate that the model has been regularized, potentially mitigating overfitting issues.
3. **Feature Importance:** While Ridge Regression doesn't perform explicit feature selection, it does shrink the coefficients of less important features. Further analysis of these coefficients could help identify the most influential factors affecting STEM degree completion rates.

Practical Implications

The high predictive power of our model suggests that universities and policymakers could potentially use similar models to forecast STEM degree completion rates with a high degree of accuracy. This could be valuable for resource allocation, program planning, and developing targeted interventions to improve STEM degree completion rates.

Limitations

1. The extremely high R-squared value is unusual in social science research. This could indicate that: a) Our model captures nearly all relevant factors influencing STEM degree completion. b) There might be some inherent structure or dependency in the IPEDS data that leads to this high explanatory power. c) There could still be some degree of overfitting, despite the use of Ridge Regression.

2. Our study is limited by the features available in the IPEDS Completions dataset. While comprehensive, this dataset may not capture all factors influencing STEM degree completion, such as individual student characteristics, institutional support programs, or local economic conditions.
3. The analysis is based on data from a single year (2022), which doesn't account for potential temporal variations or long-term trends in STEM degree completion rates.

Future Research

To build upon this study, future research could:

1. Apply other regularization techniques like Lasso or Elastic Net for comparison with Ridge Regression results.
2. Incorporate data from multiple years to examine temporal trends and improve generalizability.
3. Include additional external factors not captured in the IPEDS data to potentially uncover any missing influential variables.
4. Conduct a more detailed analysis of the most significant features identified by the model to provide targeted recommendations for improving STEM degree completion rates.
5. Perform out-of-sample validation or cross-validation to further assess the model's generalizability.

Conclusion

Our analysis provides a robust model for predicting STEM degree completion rates at public universities in the United States. While the model shows exceptionally high predictive power,

further research and validation could provide additional insights and confirm the generalizability of these findings. The results of this study can serve as a valuable starting point for policymakers and educational institutions in their efforts to improve STEM degree completion rates.

References

1. Astin, A. W. (1993). *What Matters in College? Four Critical Years Revisited*. Jossey-Bass.
2. Chen, X., & Soldner, M. (2013). *STEM Attrition: College Students' Paths Into and Out of STEM Fields* (NCES 2014-001). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from https://nces.ed.gov/pubs2014/2014001rev.pdf
3. National Academies of Sciences, Engineering, and Medicine. (2016). *Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways*. The National Academies Press. https://doi.org/10.17226/21739
4. National Science Foundation. (2020). *The State of U.S. Science and Engineering 2020*. National Science Board. Retrieved from https://ncses.nsf.gov/pubs/nsb20201/
5. Pascarella, E. T., & Terenzini, P. T. (2005). *How College Affects Students: A Third Decade of Research* (Vol. 2). Jossey-Bass.
6. Tinto, V. (1993). *Leaving College: Rethinking the Causes and Cures of Student Attrition* (2nd ed.). University of Chicago Press.

Additional Resources

- Integrated Postsecondary Education Data System (IPEDS). National Center for Education Statistics. https://nces.ed.gov/ipeds/