Predicting Student Success and Financial Outcomes in Higher Education

Aneesh Ojha
A69032336

Chutian Gong
A69041200

Kuntal Kokate
A69041623

Yiting Wang
A69044563

Zheng Dong
A69044423

1 Problem Description

Higher education represents one of the most significant financial investments individuals make in their lifetime. However, prospective students and their families often lack clear, data-driven insights when choosing colleges. Key questions remain unanswered: Which institutions provide the best return on investment? What factors predict student completion rates? How do different college characteristics affect graduate earnings and student debt levels?

This project aims to address these critical questions by analyzing the U.S. Department of Education's College Scorecard dataset. We will explore relationships between institutional characteristics (such as location, size, admission rates, and costs) and student outcomes (including graduation rates, median earnings, and debt levels). Through data analysis and visualization, we can help prospective students make more informed decisions about college selection and identify factors that contribute to positive student outcomes.

2 Dataset Overview

Source: **U.S. Department of Education College Scorecard**
URL: https://collegescorecard.ed.gov/data

The College Scorecard dataset contains comprehensive information on over 6,000 U.S. postsecondary institutions from 1996 to present, integrating data from multiple federal sources:

- **IPEDS**: Institutional characteristics, enrollment, completion rates, costs
- **NSLDS**: Federal student aid data, loan amounts, repayment status
- **IRS Tax Records**: De-identified earnings data (using differential privacy)

Key variables include institutional type and control, admission rates, tuition and net price by income quintile, completion rates (overall and by demographics), median earnings at 6/8/10 years post-enrollment, median debt levels, and loan repayment rates. The dataset includes privacy-protected data with suppression thresholds (minimum 30 students) and differential privacy techniques for earnings metrics.

3 Proposed Method

We'll use pandas and numpy to load and clean the data, handling missing values and privacy-suppressed entries. Since the dataset spans multiple years and includes various institution types, we'll need to carefully merge and filter the data to focus on meaningful comparisons.

Our analysis will explore relationships between institutional features and student outcomes. We'll compare performance across institution types and examine how outcomes vary by student demographics, particularly across income quintiles. Statistical tests will help determine whether observed differences are significant, and we'll investigate trends in costs and outcomes over the past two decades.

Following data preprocessing, we will conduct an EDA to visualize correlations between institutional costs, admission rates, and student outcomes. To build our predictive model, we will first implement a **Multiple Linear Regression** to predict continuous outcomes, such as median earnings and student debt levels. We will then utilize a **Random Forest Regression** to capture more complex, non-linear relationships and identify the key institutional factors that most significantly contribute to student success.

References

[1] U.S. Department of Education. (2024). College Scorecard Data Documentation. Retrieved from https://collegescorecard.ed.gov/data/.

[2] Bastedo, M. N., & Jaquette, O. (2011). Running in place: Low-income students and the dynamics of higher education stratification. Educational Evaluation and Policy Analysis, 33(3), 318-339.

[3] Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2020). Income segregation and intergenerational mobility across colleges in the United States. The Quarterly Journal of Economics, 135(3), 1567-1633.

[4] Dale, S. B., & Krueger, A. B. (2014). Estimating the effects of college characteristics over the career using administrative earnings data. Journal of Human Resources, 49(2), 323-358.