Quantifying Socioeconomic Disparities in Post-Graduate Earnings: A Machine Learning Approach Using SHAP

Ansh Singh

July 2025

1 Abstract

In this study, I investigated the relationship between college characteristics and post-graduate earnings among US students who received Pell grants and non-recipients. Data were taken from Department of Education College Scorecard dataset. I trained a machine learning model to predict income after 5 years of graduation, separately, for both types of students. A Random Forest Regressor achieved R^2 scores of 0.71 and 0.70 for Pell and non-Pell respectively, indicating high accuracy. To understand and explain the key factors influencing income disparity, I used SHAP (SHapley Additive exPlanations). My analysis shows the most influential institutional and academic factors which contributed to income inequality, such as program type, program duration, degree level, and institutional control. I also trained a model to directly predict the income gaps between the two groups, achieving RMSE of $9508.

2 Introduction

Access to higher education is often seen as a pathway to economic stability and financial freedom. However, the extent to which higher education reduces or reinforces socioeconomic disparities remains debated. One metric that can shed light on this issue is post-graduate earnings. Pell Grants are awarded to low-income students in the US, and examining their earnings relative to their non-Pell counterparts offers a glimpse into persistent inequality. This paper seeks to quantify and explain income disparities between Pell and non-Pell students using machine learning and interpretable AI techniques.

3 Dataset and Preprocessing

3.1 Features

The selected features included:

- CONTROL: Institutional control (public, private nonprofit, private for-profit)
- CIPDESC: Program/major description
- CREDDESC: Credential description (e.g., Bachelor's Degree)
- CREDLEV: Credential level (numeric)
- EARN_PELL_WNE_MDN_5YR: Median income for Pell Grant students 5 years post-graduation
- EARN_NOPELL_WNE_MDN_5YR: Median income for non-Pell students 5 years post-graduation

3.2 Preprocessing

- Replaced all "PrivacySuppressed" values with NA and dropped rows with missing earnings data.
- Categorical variables (CONTROL, CIPDESC, CREDDESC) were one-hot encoded. Also, I kept all the categories in **CONTROL**(you will know what I'm talking about in section 5.2).Therefore, there is no baseline. All categories are treated independently, and the model will learn from all of them (which can cause multicollinearity — but not an issue in this case, since I used RandomForest).
- Target variables (EARN_NOPELL_WNE_MDN_5YR, EARN_PELL_WNE_MDN_5YR) were converted to numeric types for regression.

4 Methodology

I trained two Random Forest Regressor models:

1. Model 1: Predict Pell student income
2. Model 2: Predict non-Pell student income

then calculated the difference between predicted incomes to assess the income gap. A third model was trained directly to predict the gap (Model 3).

5 Results

5.1 Performance Metrics

The first thing you may notice is the RMSE score is off by approx 15k-16k USD, considering about 22-24% relative error and for the real world, it isn't that bad. I'm saying this because only institutional & degree-level data is being used for the prediction, no **SAT scores, GPA, gender, race, work experience**, etc. Similarly, the low R^2 score of the gap model suggests that college/program characteristics alone do not explain the income disparity, hence the importance of other factors I mentioned above. Here are the performance metrics of the different models trained:

Model	**R^2**	**RMSE (USD)**	**Mean (USD)**	**MAE (USD**
Pell Model	0.7197	15,851	66,082	9,400
Non-Pell Model	0.7026	16,266	71,220	9,914
Gap Model	0.1305	9,508	–	6,459
Income Gap (Predicted)	–	–	5,138	–

Table 1: Performance Metrics of the Models

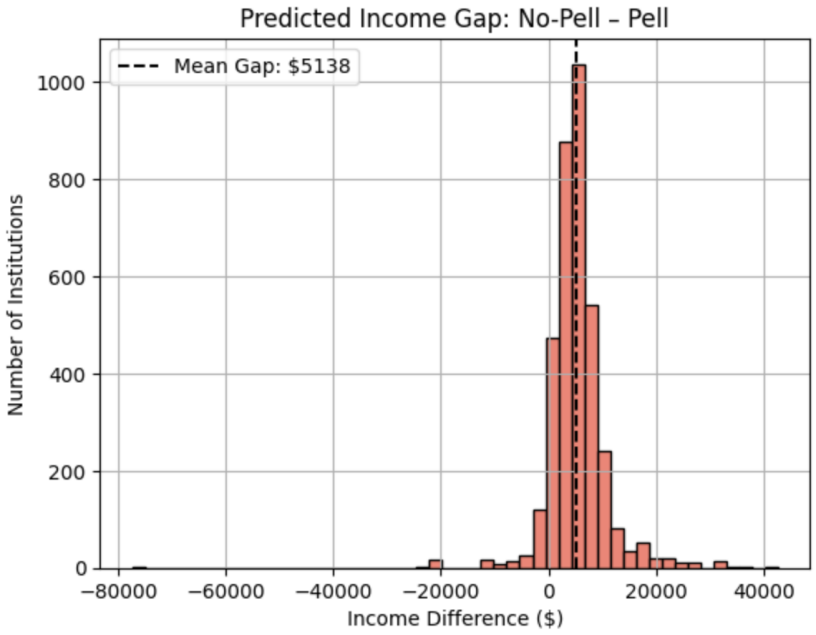

Figure 1: Income Gap Prediction (Model 3)

5.2 Feature Importance via SHAP

SHAP was used to interpret the contributions of each feature. To sum up, the summary plot revealed that institutional control played a critical role in explaining income disparity. Specifically, **for-profit private** institutions was the strongest positive contributor to the income gap, which means these institutions disadvantage Pell Grant recipients or even normal students. In contrast, **private non-profit** institutions showed a lower or even negative SHAP contribution, indicating a more equal outcome across socioeconomic backgrounds. Public universities had a slightly increasing gap.

- **Credential Level (CREDLEV)**: Higher degree levels (e.g., Ph.D.) had a strong positive impact on predicted income.
- **Field of Study (CIPDESC)**: Business, Computer Science, and Engineering were among the top contributors to higher incomes.
- **Institution Type (CONTROL)**: Private non-profit universities had a positive impact; for-profit institutions contributed negatively.
- **Credential Description (CREDDESC)**: Bachelor's and First Professional degrees contributed positively.

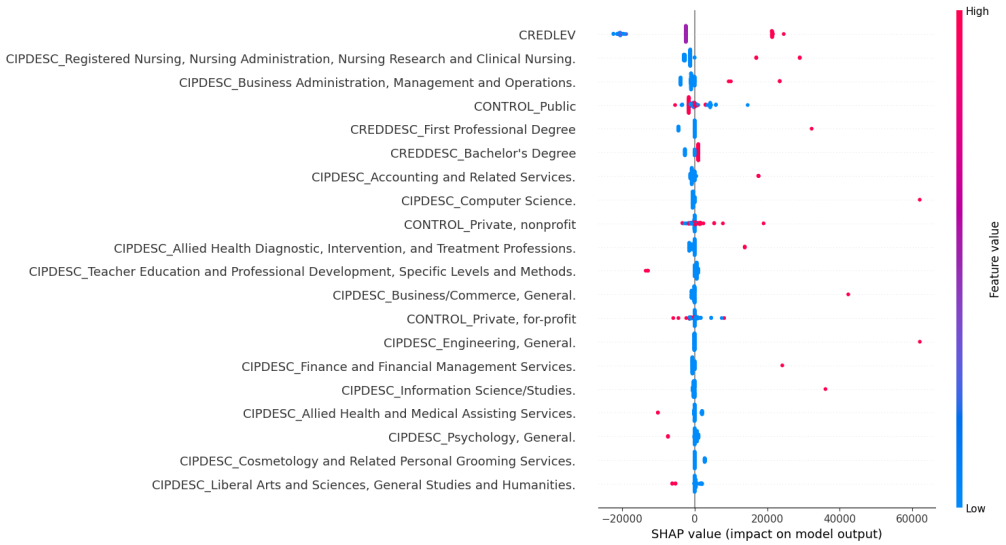

Figure 2: SHAP values of Income Prediction—Pell Grant Students (Model 1)

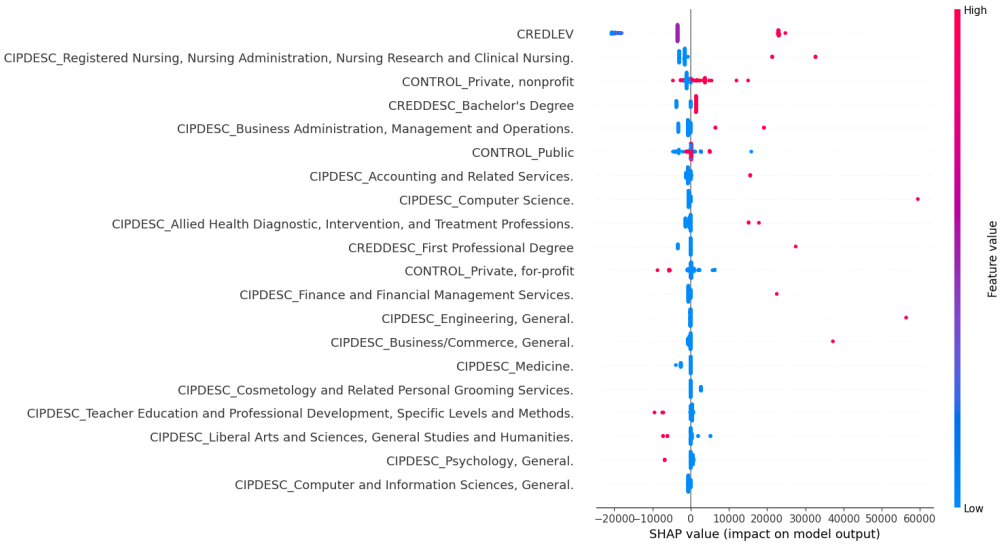

Figure 3: SHAP values of Income Prediction—NoPell Students (Model 2)

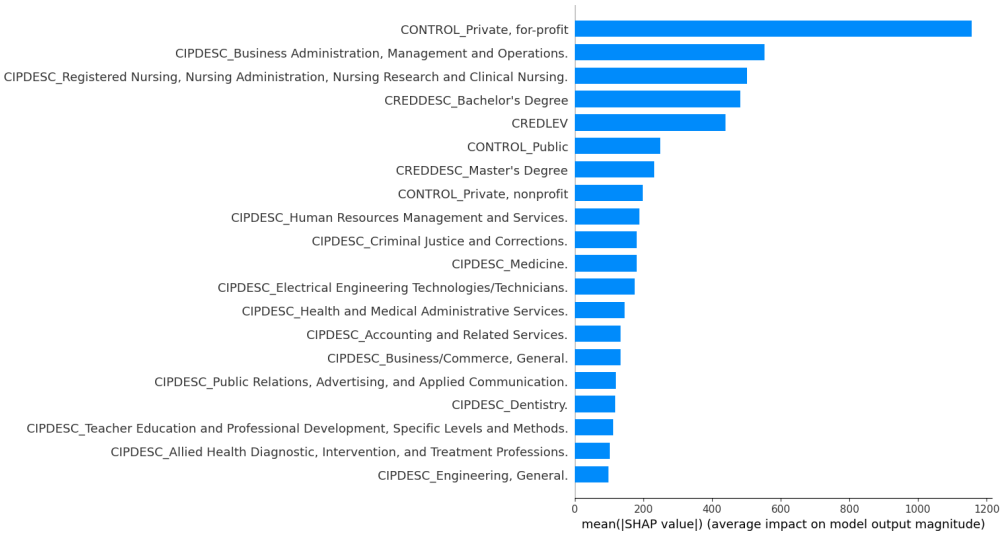

Figure 4: SHAP values of Income Gap (Model 3)

6 Discussion

The results show that not all degrees and institutions provide equal ROI. Even students who are doing similar degrees, institutional factors matter significantly more. Fields such as Engineering and Computer Science offer high returns, while degrees in Liberal Arts, Psychology, etc. tend to predict lower earnings.

The $5,138 income gap between Pell and non-Pell students suggests that socioeconomic disparities persist even after accounting for degree and institution. This may reflect unmeasured variables such as **networking opportunities, access to internships**, discrimination in the job market and other factors including individual differences, work ethic, cognitive ability and perseverance.

7 Acknowledgments

Thanks to the U.S. Department of Education for making the data openly accessible.

8 References

- U.S. Department of Education College Scorecard: https://collegescorecard.ed.gov/data/
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.