Title: Predicting Engineering Graduate Earnings Using Institutional and Program-Level Data
Authors: Elias Krasny, Jiei Ota, William Wright, Elias Zell
Course: SYS 3501
Date: December 19, 2025
Repository: https://github.com/eliaskrasny/SYS3501_Final-Engineering-Earnings-Analysis

Executive Summary

This project develops and evaluates a comprehensive data science pipeline designed to predict post-graduation earnings for engineering programs in the United States. Using large-scale institutional and program-level data from the U.S. Department of Education, the analysis integrates data engineering, exploratory data analysis, feature engineering, regression modeling, and classification techniques to understand and predict earnings outcomes three years after graduation. The core motivation is practical: students, policymakers, and institutions all want to understand what drives earnings differences across engineering programs, and whether those differences can be predicted accurately and explained meaningfully.

The project addresses two central problems. First, can we accurately predict median earnings for engineering graduates using observable institutional and demographic characteristics? Second, can we reliably classify engineering programs into intuitive earnings tiers (low, medium, high, very high) that are more actionable for decision-making than precise dollar predictions alone? To answer these questions, we constructed a dataset of 8,880 engineering programs, engineered theoretically motivated features, and evaluated multiple modeling approaches ranging from highly interpretable linear regression to high-performance ensemble tree methods.

The results are decisive. Linear models explain roughly 74–75 percent of the variance in earnings and provide clear, interpretable “weighted sum” relationships between institutional features and outcomes. In contrast, Random Forest and Gradient Boosting models achieve near-perfect predictive performance, explaining over 98 percent of the variance and classifying programs into earnings buckets with nearly 98 percent accuracy. Across all analyses, out-of-state tuition emerges as the single dominant predictor, acting as a powerful proxy for institutional prestige, resources, and student quality. Faculty salary, completion rates, and selectivity metrics also play substantial roles.

This project demonstrates both the technical rigor of a full data science pipeline and the substantive insight that institutional context matters enormously for early-career engineering earnings. The findings have direct implications for students choosing programs, institutions allocating resources, and researchers studying inequality and returns to higher education.

1. Research Questions and Objectives

The analysis is guided by two primary research questions. First, to what extent can engineering graduate earnings three years after graduation be predicted using institutional characteristics, program attributes, and student demographics? Second, which factors matter most in explaining differences in earnings across engineering programs, and how do those factors interact?

From these questions flow several concrete objectives. The first objective is to build a clean, integrated dataset linking engineering programs to their institutional characteristics. The second objective is to explore the distribution of earnings and identify key correlates through exploratory data analysis. The third objective is to engineer new features that capture theoretically meaningful interactions and non-linearities. The fourth objective is to build and evaluate multiple predictive models, balancing interpretability and accuracy. Finally, the project aims to translate technical results into practical insights that can inform real-world decision-making.

2. Data Description

The data used in this project come from the U.S. Department of Education's publicly available College Scorecard datasets. Two primary files form the backbone of the analysis. The first is the Field of Study dataset, a 142 MB CSV file containing 229,188 records. Each record represents a specific academic program at a specific institution, such as Mechanical Engineering at MIT or Civil Engineering at Virginia Tech. The second is the Institution dataset, a 98 MB CSV file containing 6,429 records, one for each college or university, with over 1,000 variables describing institutional characteristics such as tuition, selectivity, demographics, spending, and faculty salaries.

The scope of this project is restricted to engineering programs. Engineering programs are identified using Classification of Instructional Programs (CIP) codes ranging from 1400 to 1499. Examples include 1408 for Civil Engineering, 1410 for Electrical Engineering, and 1419 for Mechanical Engineering. Filtering on these codes reduced the dataset from 229,188 total programs to 8,880 engineering programs.

The two datasets were merged using the UNITID variable, a unique institutional identifier present in both files. A left join was performed to retain all engineering programs while appending institutional characteristics. The result was a merged dataset with 8,880 rows and 3,479 columns, representing nearly 30 million individual data points. This scale necessitated careful data cleaning, feature selection, and computational efficiency.

Key variables in the analysis include median earnings three years after graduation (the primary outcome variable), tuition levels, faculty salary, student demographics, SAT and ACT scores, acceptance rates, completion rates, and measures of institutional spending. These variables were selected based on prior research, domain knowledge, and empirical relevance.

3. Exploratory Data Analysis

The outcome variable for this project is median earnings three years after graduation, measured by the variable MD_EARN_WNE_INC3_P7. An initial examination of its distribution reveals substantial variation across engineering programs. The minimum observed median earnings are $24,920, while the maximum reaches $140,193. The 25th percentile is $47,044, the median is $54,679, and the 75th percentile is $65,766. The mean is approximately $56,500, with a standard deviation of about $15,400.

The distribution is right-skewed, with a long tail of very high-earning programs. Most engineering programs cluster between $45,000 and $70,000, but the wide overall range indicates that institutional context plays a major role. Even within a field that is generally well-paid, where the median exceeds $54,000 just three years after graduation, differences of tens of thousands of dollars are common.

Several key predictors were explored in depth. Out-of-state tuition ranges from roughly $5,000 to $65,000, with a mean near $32,000. This variable shows an exceptionally strong relationship with earnings and ultimately becomes the single most important predictor in the models. Faculty salary ranges from about $40,000 to $150,000, with a mean around $78,000, and correlates strongly with earnings. SAT scores range from 800 to 1550, with a mean near 1150, and also show a very strong positive association. Acceptance rates vary from highly selective institutions admitting 5 percent of applicants to open-access institutions admitting nearly all applicants, with a median acceptance rate around 65 percent. Acceptance rate is negatively correlated with earnings, reflecting the role of selectivity.

A comprehensive correlation analysis was conducted between earnings and all candidate predictors. The strongest positive correlations include standardized test scores, out-of-state tuition, faculty salary, and completion rates. The strongest negative correlations include acceptance rate, percentage of Pell Grant recipients, and percentage of Black students. All of these correlations are statistically significant at $p < 0.001$, reflecting the enormous statistical power afforded by more than 8,000 observations. These results do not imply causation, but they clearly identify which variables are most strongly associated with earnings outcomes and therefore warrant deeper modeling.

4. Methodology

Data preparation began with extensive cleaning to address privacy suppression and missing values. Federal privacy rules suppress data for programs with fewer than ten students, resulting in placeholders such as “PrivacySuppressed,” “NULL,” or empty strings. All such indicators were systematically replaced with proper missing value markers, and columns were converted to numeric types wherever possible.

Missing data posed a major challenge. Approximately 8 percent of programs were missing earnings data, while debt variables were missing for roughly 80 percent of programs. Many institutional variables had between 10 and 50 percent missingness. For earnings modeling, features with more than 50 percent missing values were removed. Remaining missing values were imputed using medians, a choice motivated by robustness to outliers. SAT scores with missing values were filled with the median SAT score of approximately 1150. Debt outcomes were analyzed separately due to their extreme missingness. After filtering, 8,230 programs had complete earnings data and were retained for modeling.

Feature engineering played a central role. Ten new features were created, including interaction terms, polynomial terms, ratio variables, and binary indicators. Examples include interactions between selectivity and resources, quadratic terms for SAT and acceptance rate to capture non-linear effects, ratios measuring resource allocation efficiency per Pell Grant student, and indicators for institutional type, urban location, and elite status. Several candidate features were explicitly rejected due to overfitting risk, multicollinearity, weak empirical relevance, or ethical concerns.

For modeling, the data were split into training and test sets using an 80/20 split with a fixed random seed to ensure reproducibility. Features were standardized using z-score normalization for models sensitive to scale, such as Ridge and Lasso regression.

Four regression models were evaluated. Ordinary least squares linear regression provided a baseline and yielded an R-squared of 0.745, with a mean absolute error of about $5,565. Ridge regression produced nearly identical performance, indicating that multicollinearity was not a major problem. Lasso regression slightly reduced performance but selected a subset of 16 important features, providing insight into which variables truly matter. Random Forest regression dramatically outperformed all linear models, achieving an R-squared of 0.983 and a mean absolute error of only $571.

In addition to regression, classification models were built to predict earnings buckets defined by quartiles. Logistic regression achieved modest accuracy of about 69 percent. Random Forest and Gradient Boosting classifiers achieved near-perfect performance, with accuracies of 97.7 percent and 97.9 percent respectively, and almost no severe misclassifications.

5. Results and Discussion

The most striking result is the dominance of out-of-state tuition as a predictor. In the Random Forest model, it accounts for nearly 60 percent of total feature importance. Across models, tuition acts as a super-proxy for institutional prestige, resources, student quality, and alumni networks. Faculty salary also exhibits a large marginal association with earnings, reinforcing the importance of institutional investment. Selectivity shows diminishing returns, as revealed by the negative coefficient on the squared SAT term, indicating that gains from higher selectivity taper off at very high levels.

Engineered features added meaningful value, improving predictive accuracy and uncovering non-linear patterns that would otherwise remain hidden. Classification models demonstrated that programs can be placed into intuitive earnings tiers with extremely high confidence, providing actionable information for students.

The trade-off between interpretability and accuracy is clear. Linear models are easier to explain and useful for policy discussions, while tree-based models are superior for prediction. Both are necessary, depending on the question being asked.

6. Future Work and Reflection

Future work could incorporate additional years of earnings data, alternative outcome measures, or quasi-experimental designs to better address causality. More granular program-level data and labor market controls could further refine predictions.

This project faced challenges related to missing data, feature selection, and computational scale, all of which were addressed through deliberate methodological choices. The use of modern machine learning tools, combined with domain knowledge and careful interpretation, proved highly effective.

References

1. U.S. Department of Education College Scorecard
   https://collegescorecard.ed.gov/data/