Institutional Bias in Post-Secondary Education and its Predictive Power with Respect to Post-Graduation Earnings

Jesse Swanson (js11133), Khasi Jamieson (kmj426), Will Egan (wve204)

Abstract

Post-secondary education is a trillion dollar industry in the United States. Baseline tuitions are tens of thousands of dollars and post-secondary education is a tremendous expense for those who undertake it. This financial strain is coupled with widespread misinformation about expected earnings post-graduation. The purpose of this report is to shed light on these confounding factors by using data-driven insights to estimate the earnings of graduates from post-secondary institutions and to explain the factors impacting these earnings.

Business Understanding

Differences between realities and expectations of post-graduation earnings cause prospective students to make uninformed decisions about post-secondary education. To alleviate the strain of misinformation among prospective students, our goal is two-fold:

1. Determine important attributes of post-secondary institutions resulting in high median earnings for graduates
2. Predict median earnings after graduation based on the current characteristics of institutions. This model could predict earnings six years ahead of reported earnings data (a unique competitive advantage in college rankings).

Under the assumption that US society is inherently biased against non-white, poor, female, and other traditionally marginalized populations, we hypothesize that impactful predictive features that are related to structural injustice. We hope that such findings can be presented as data-driven evidence that such issues are chief indicators of financial success. We hope that these findings inform prospective students about their expected earnings after graduation and the factors affecting their prospective earnings. Furthermore, degree-granting institutions could be held accountable for unfair inconsistencies in outcomes of students. Predicted earnings generated from machine learning (ML) models could be a comparison metric for colleges or an input for projected debt repayment models.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

Data Understanding

Data Overview

The United States Department of Education's College Scorecard data set is composed of yearly federal reports from approximately 6800 post secondary institutions from academic years 1996-1997 to 2017-2018. Data is reported to and fetched from three primary sources: the Integrated Postsecondary Education Data System (IPEDS), the National Student Loan Data System (NSLDS) and Administrative Earnings Data from the IRS data [2].

Institution Data

IPEDS records data from surveys administered by the Department of Education's National Center for Education Statistics (NCES) to all Title IV post-secondary institutions. Title IV institutions are schools that process federal aid. The surveys record school specific information on tuition, graduation rates, retention rates, cost of attendance, school size, demographic information and faculty salaries. Some non-Title IV institutions that do not participate in the program but are reported to have similar characteristics as Title IV, institutions are included in the data set. [2]

Aid Data

Federal aid data referring to federal student loans and pell grants is obtained from the National Student Loan Data System (NSLDS). NSLDS data tracks federal grant and loan disbursements, students' repayments on loan balances, borrowing status, and enrollment status of aid recipients. College Scorecard reports that approximately 70 percent of all post secondary students are federally aided. Approximately 80 percent of students of for-profit institutions receive federal aid and within non-profit institutions, 77 percent at private institutions, 59 at public 4-year institutions, and 33 at 2 year institutions. [2]

Earnings Data

Post graduation earnings data of students that receive federal aid is retrieved from tax records maintained by the United States Department of the Treasury. Income metrics are represented by mean and median earnings from W-2 forms plus reported self-employment earnings from Schedule K forms. Earnings data is captured by mean and median earnings of workers 6 to 10 years after first enrolling in the institution. Thus, earnings data for a specific year represents non-enrolled students who first-enrolled at the institution 6 to 10 year prior

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

to the year of the data set. Earnings of non-enrolled students are recorded regardless of completion status with the exception of students who continue to graduate studies. Earnings data is suppressed for graduating cohorts of less than 50 students. [2]

Overview of Features

The features in the data set are split into 9 different categories.

- Academics (272 features)
- Aid (40 features)
- Completion (1218 features)
- Cost (77 features)
- Earnings (76 features)
- Repayment (132 features)
- Root (5 features)
- School (175 features)
- Student (113 features)

Limitations of the Data

The data collected is only representative of students who received federal aid. Therefore, the data set is inherently biased towards students receiving federal aid and is limited in providing a comprehensive analysis or prediction of all post-graduate earnings. The data set may not accurately represent institutions with low percentages of students receiving federal aid.

Some information from the NSLDS and the Department of Treasury is protected to maintain student and family anonymity. Aid, repayment, and earnings categories contain features that are suppressed for privacy. Furthermore, our target variable, median earnings of graduates 6 years after entry, is likely correlated with completion rates since non-enrolled non-graduates are also included in earnings data. We cannot observe scenarios where schools have small completion rates and high earnings for graduates since earnings are reported for graduates and non-graduates.

College Scorecard provides limited insight into the reporting process and states there are potential inconsistencies in institution reporting due to the individual complexity of each institution. Institutions are commonly composed of complex systems of administration that will vary from school to school. Without further specifications of reporting process, we must accept this inconsistency as an irreducible error within reported data.

The data set contains feature naming inconsistencies across years. For example, due to updated reporting practices in the US census, racial demographic classifications may change. College Scorecard Data introduces new features that are representative of the new classifications

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

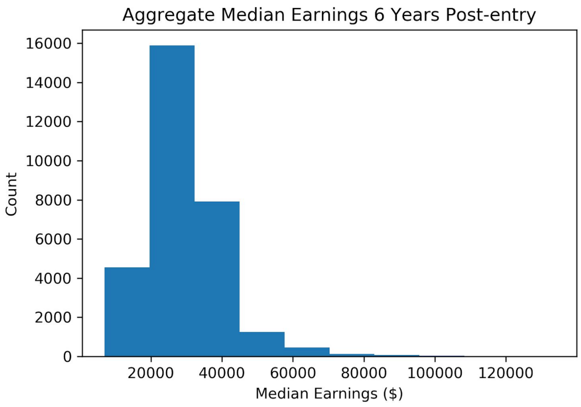
and no longer populates the older related features.

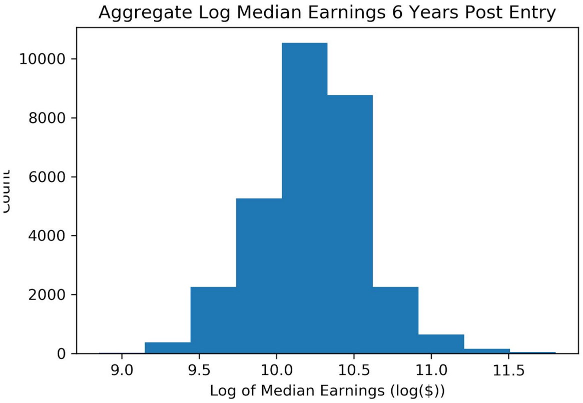
Data Preparation

The College Scorecard data set contains 1977 features. To reduce the number of features used in the ML model, duplicate features and uninterpretable features were removed from the data set. For example, degrees issued by major were separated into two classes, amount issued ‘CIP’ and share of degrees issues ‘PCIP’. Only one group of features is necessary since they are related to the size of the institution. Features containing leakage with respect to the target variable such as other earnings data were removed.

Target

The target variable for our ML model is the median earnings of students who are employed and no longer enrolled in a post secondary institution 6 years after entry. Earnings 6 years after entry was chosen as the target variable since federal Pell grants can only be received for 12 semesters or approximately 6 years after entry. Furthermore, earnings 6 years after entry provided the most data points for the ML model. For a given year, this information is recorded by the MD_EARN_WNE_P6 feature. The median is immune to outliers and more representative of earnings graduates can expect when compared to the mean. The target variable is positively skewed and is log transformed for analysis.





Data Cleaning and Feature Engineering

The College Scorecard data set is large and sparse. Unpopulated features have low variance and will have less impact on ML model predictions than features with more variance. Therefore, features with more than 90% missing values were dropped. Remaining missing values for non-categorical features were replaced with the mean of the feature. Instances missing the

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

target variable were dropped since the model cannot be trained on an instance with no target information and predictions cannot be validated.

Over 50 encoded features represented shares of students in different fields of study at post-secondary institutions. To increase the significance of these features and to increase the interpretability of our model, the features were binned into the following overarching fields of study: Natural Sciences, Engineering, Arts, Professional Studies, Humanities, Law, Medical, Military, and Business.

To ensure the ML model is school-agnostic, all identification numbers were dropped to prevent overfitting. Furthermore, regression ML models cannot predict using categorical variables and these variables were one-hot encoded. The magnitudes of the remaining features varied and large values could dominate linear models. To compensate for this, all non-flag features were standardized.

To determine if linear ML models would be susceptible to multicollinearity, a correlation matrix was generated using the top 20 features as measured by F-tests in regression. Based on this analysis, family income, first generation student percentages, and independant income had duplicate features. Duplicate features were dropped based on this analysis.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

Modeling & Evaluation

Evaluation Metric

Mean-squared error (MSE) and mean-absolute error (MAE) are common metrics for evaluating regression models. MSE is the model evaluation metric for the following analysis as it penalizes large prediction errors more than MAE. However, MAE is more interpretable and it has been calculated for baseline and final model comparisons. ML models will be compared using the MSE of the natural log of income.

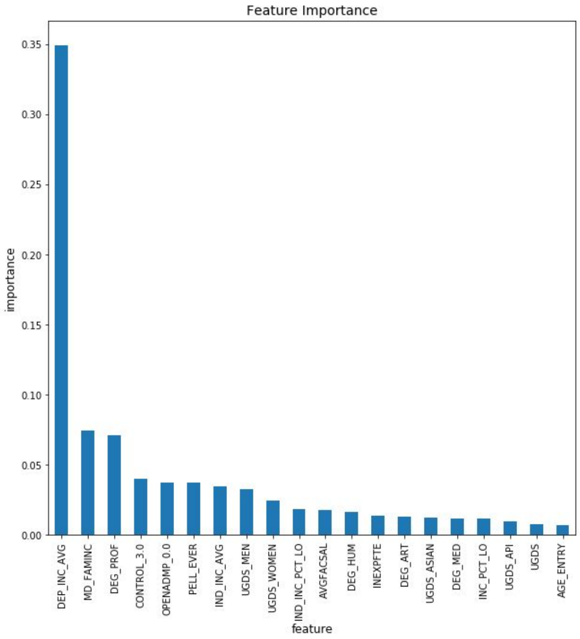
Baseline Model

The baseline model is a simple year over year wage growth model based on an average 3% wage growth in the United States. Median earnings for the test data sets of 2013 and 2014 were generated based off median earnings in 2012, the latest training set year. Various spline extrapolation models to generate median earnings for 2013 and 2014 were considered but produced worse MSE than the basic wage growth model. This model was chosen as a baseline since future post-graduation earnings may be best represented by historical earnings of an institution. Subsequent machine learning models attempt to outperform the simple baseline model. The baseline model has a MSE of 0.0049 (MSE is calculated based on natural log of earnings) and MAE of $1313 (MAE is calculated based on earnings for interpretability).

Preliminary Random Forest Evaluation

A default Random Forest Regressor with 20 estimators (trees in the forest) was trained using the 187 features remaining after our feature engineering. The goal of this model is two-fold: to determine the highest feature importance based on the greedy approach of the algorithm and to obtain a preliminary MSE of a random forest model. The model has an MSE of 0.0018.

We were concerned that the random forest model with 187 features might have overfit to our training set. Furthermore, a model with 187 features lacks interpretability and contains features with overlapping mutual information gain. To address these concerns, we limit the number of features to 20 in future models. This choice is further supported by the small values of feature importance beyond the top 20 features. Reducing the number of features increases the bias of the model and reduces the variance to protect the model from overfitting the data.



The default random forest was run with the top 20 features determined by the previous model. The default random forest with 20 features has a MSE of 0.0020 (MSE is calculated based on natural log of earnings) Since this is lower than the initial baseline MSE of 0.0049, 0.0020 will be the baseline for subsequent models.

Modeling

To predict and explain features important to the target variable, the following models were used as specified in scikit-learn: Random Forest Regressor, SVM Regressor, Linear Regression, and Lasso regression. The dataset is split into training (75%) and test sets (25%) based on

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

year. The training set is composed of data for years 2003-2012 (6 data points per institution based on College Scorecard reporting due to binning years) and the test set is composed of data for years 2013-2014 (2 data points per institution). The initial RF model was evaluated using the same training/test split.

Given the large number of features remaining (187), three feature selection methods were employed to improve the interpretability of the model and reduce the number of features to 20. These methods will return the most important features for prediction based on their respective metrics. The three feature selection methods used were univariate selection using F-tests, mutual information and recursive feature elimination with a random forest. Each feature selection method was run with each of the four models and evaluated based on MSE. The model with the lowest MSE will be chosen as the optimal feature/model combination.

Random Forest Regressor (RF)

Random forests perform well with high dimensional data sets and non-linear data Additionally, RFs perform well with sparse data sets, a necessity for the sparse College Scorecard data set.

Support Vector Regressor (SVR)

Support vector regressors support non-linear relationships between features and the target variable with a non-linear kernel. However, the datasets must be linearly separable which may not be the case in the College Scorecard data set.

Linear Regression (LIN)

Linear regression is an effective model when relationships between features and the target variable are strictly linear. However, the College Scorecard data may not be linear thus we expect a high error rate .

Lasso Regression (LASS)

Initial RF results imply a few features dominate predictions of the target variable. Lasso regression model forces feature coefficients to zero when they are not important and improves accuracy when only a few features dominate. Therefore, Lasso regression performs regularization when fitting the data resulting in better accuracy and lower bias when compared to linear regression. However, Lasso regression may suffer from the same drawbacks as linear regression.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

Feature Selection Tools

F-Test in Regression (FR)

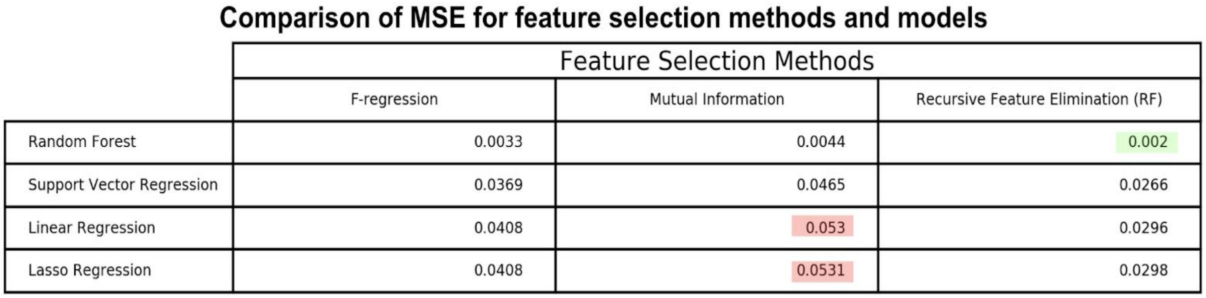
This feature selection method is based on F-Tests for regression. It calculates the correlation between each feature in the training set and the target variable, which it then converts into an F score and then a p-value. We rank the variables based on these values, and then select the 20 best.

Mutual Information Regression (MIR)

MIR estimates the mutual information between each feature in the training set and the target variable. Similar to the FR case, we rank our features based on these values and pick the 20 highest.

Recursive Feature Elimination (RFE)

RFE using a RF trains a RF and marks the least significant feature for removal. The feature selection recursively removes the weakest features until the specified number is reached. The goal of this feature selection method is to complement the greedy mutual information gain approach taken by RF by selecting features from the bottom up instead.



Comparison of MSE for feature selection methods and models

	Feature Selection Methods		
	F-regression	Mutual Information	Recursive Feature Elimination (RF)
Random Forest	0.0033	0.0044	0.002
Support Vector Regression	0.0369	0.0465	0.0266
Linear Regression	0.0408	0.053	0.0296
Lasso Regression	0.0408	0.0531	0.0298

Recursive feature elimination using a random forest resulted in the lowest mean-squared error. Lasso regression and linear regression using mutual information to determine the top 20 features for modeling has the highest mean-squared error. Mutual information feature elimination captures non-linear relationships between features and the target variable. Therefore, it is expected that features chosen using mutual information will result in poor fits for linear models.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

Evaluation

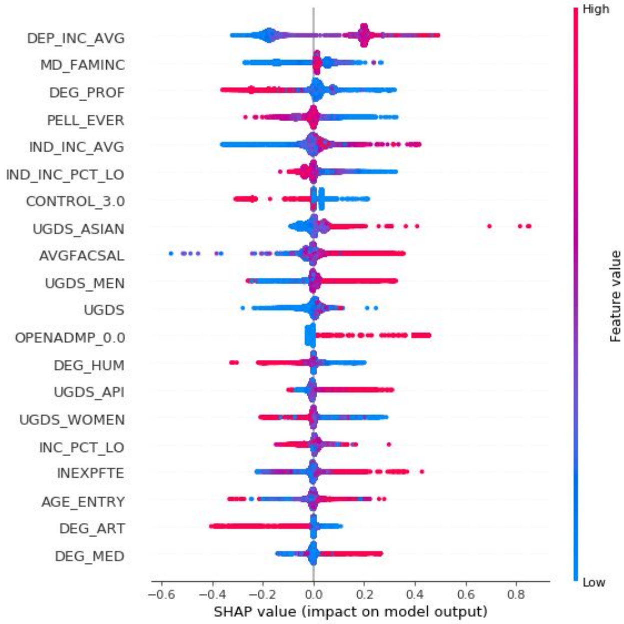
Each ML model was trained on a dataset outputted by the previous feature selection methods. Each model and feature selection combination was evaluated based on the MSE of target predictions against the test values. The Random Forest with RFE selected features has the lowest MSE.

To tune the RF model, we perform a 5-fold cross validation across 20, 30, 50, 100, 150, and 200 estimators (number of trees). The RF has the lowest MSE when using 150 estimators. The final evaluation metrics for the RF are a MSE of 0.0016 and a MAE of $887. This model is an improvement over both the initial random forest with 187 features (MSE score of 0.0020 and a MAE of $1313) and the baseline year over year wage growth model (MSE of 0.0049 and MAE of $1313).

SHAP Plot

SHAP is a visualization tool used to explain model predictions. The SHAP diagram plot how each feature contributes to the predictions of the RF.

SHAP Plot for RF with Income Features



In the above plot, dots represent instances of the data set. If a dot is red, the corresponding feature has a high value and blue corresponds to a low value. The SHAP values are proportional to the impact of the feature on the target variable value. For example, DEG_MED has red point with a SHAP value of 0.2. For this particular instance, a high share of students seeking medical degrees results in a higher target variable prediction. This makes intuitive sense, since doctors are among the highest paid professionals. Conversely, high shares of students seeking degrees in the humanities or the arts (captured by DEG_HUM and DEG_ART, respectively) are inversely proportional to median earnings.

[1]: *Using Federal Data...*2015

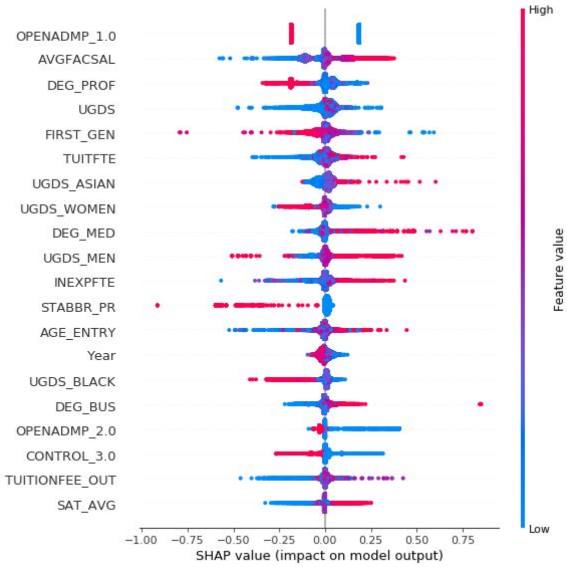
[2]: Data Dictionary

SHAP values capture potential social injustices such as high shares of female undergraduates (UGDS_WOMEN) being inversely proportional to earnings and high shares of male undergraduates (UGDS_MEN) being proportional to earnings.

Possible Leakage?

Many significant features are related to student or family income. This is a potential source of data leakage since family or student earnings before post-secondary education are related to earnings after graduation. The RF model with RFE feature selection is run again but excluding features related to financial aid and income. This RF has a MSE of 0.0025, a significant increase from the tuned RF including family income (MSE of 0.0016).

SHAP Plot for RF Without Income or Financial Aid Features



The SHAP plot suggests that the values of the following features are proportional to the target variable:

- high average SAT scores (AVG_SAT)
- high faculty salaries (AVGFACSAL)
- high instructional expenditures per full-time equivalent student (INEXPFTE)

The SHAP plot suggests structural injustice has a substantial effect on the financial success of graduates of degree-seeking institutions. The shares of the following demographics are inversely proportional to the target variable:

- high shares of black undergraduates (UGDS_BLACK)

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

- high shares of female undergraduates (UGDS_WOMEN)
- An address in Puerto Rico (STABBR_PR)

An open admissions policy (OPENADM_1.0) is an example of an institution-specific feature and is a signal of lower median earnings after graduation.

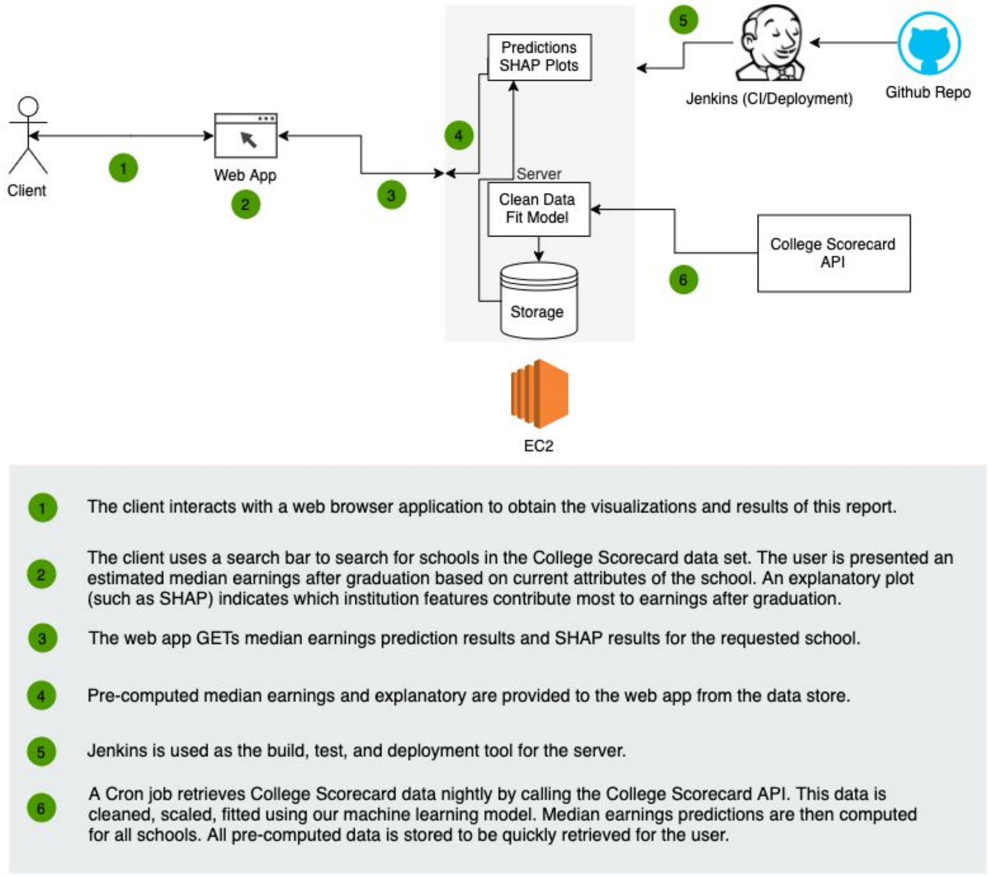
The SHAP plot accurately identifies marginalized groups across race (UGDS_BLACK), gender (UGDS_WOMEN), geographic location (STABBR_PR), and economic status (disadvantaged students have far less access to resources with which to study for the SAT and subsequently get higher scores).

Deployment

The results of this report will be deployed as a web application. Users will have the ability to search for an institution of interest and predict their median earnings six years after entry. Visualization of features of the school that contribute to the estimated earnings of the ML model will be displayed as well.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary



1 The client interacts with a web browser application to obtain the visualizations and results of this report.

2 The client uses a search bar to search for schools in the College Scorecard data set. The user is presented an estimated median earnings after graduation based on current attributes of the school. An explanatory plot (such as SHAP) indicates which institution features contribute most to earnings after graduation.

3 The web app GETs median earnings prediction results and SHAP results for the requested school.

4 Pre-computed median earnings and explanatory are provided to the web app from the data store.

5 Jenkins is used as the build, test, and deployment tool for the server.

6 A Cron job retrieves College Scorecard data nightly by calling the College Scorecard API. This data is cleaned, scaled, fitted using our machine learning model. Median earnings predictions are then computed for all schools. All pre-computed data is stored to be quickly retrieved for the user.

Risks and Ethical Concerns

The ML models in this report estimate earnings based on controversial features such as demographics. While some of our findings could have upsetting implications, the goal of our ML model is not to dissuade students from attending a school based on the school's demographics. These results are an investigation into structural injustices in the education and social system and are meant to inform users to the real effects structural issues in post-secondary education. Furthermore, prospective students have different objectives when attending post-secondary institutions that are not addressed in the scope of our project. Our results do not capture student goals such as preparation for graduate studies or quality of education.

Improvements

The College Scorecard data lacks features differentiating earnings based on faculty or program. This is a crucial insight that could be used to compare estimated earnings based on field of study. Furthermore, more work is

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

required to determine the quality of an institution. Determining the relative impact of an institution by compensating for financial aid or income attributes remains an area of open research.

Conclusion

The findings of this report support the hypothesis that institution features relating to structural injustice are effective predictors of median earnings after graduation. The most effective machine learning model tested was a Random Forest model trained on 20 features selected by Random Forest recursive feature elimination. The RF model had a MSE of 0.0017 compared to the baseline wage growth model with a MSE of 0.0049. In this model, family income was the most significant factor to determine earnings after graduation.

To address potential data leakage in this model, all income-related features were removed. The resulting Random Forest model was less accurate and had a MSE of 0.0025. Given that features encoding income information are the significant by every feature selection method, reduced accuracy is expected. Resulting SHAP plots suggest values of features related to marginalized demographics are inversely proportional to expected median earnings after graduation.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

References

[1] US Department of Education, College Scorecard. *Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education,* September 2015.

[2] US Department of Education, College Scorecard. *College Scorecard Data Dictionary*, https://collegescorecard.ed.gov/data/documentation/,

Appendix

Feature Glossary

AGE_ENTRY	Average age of entry
AVGFACSAL	Average faculty salary
CONTROL	1 = Public 2 = Private non-profit 3 = Private for-profit
DEG_ART	Share of students in arts
DEG_BUS	Share of students in business
DEG_HUM	Share of students in humanities
DEG_MED	Share of students in medicine
DEG_PROF	Share of students in professional studies
DEP_INC_AVG	Average family income of dependent students in real 2015 dollars
FIRST_GEN	Share of first generation students
INC_PCT_LO	Percentage of aided students whose family income is between $0-$30,000
IND_INC_AVG	Average family income of independent students in real 2015 dollars.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

IND_INC_PCT_LO	Percentage of students who are financially independent and have family incomes between $0-30,000
INEXPFTE	Instructional expenditures per full-time equivalent student
MD_FAMINC	Median family income in real 2015 dollars
OPENADMP	Open admissions policy indicator. 1 is yes. 0 is no.
PELL_EVER	Share of students who received a Pell Grant while in school
SAT_AVG	Average equivalent SAT score of admitted students
STABBR_PR	State = Puerto Rico
TUITFTE	Net tuition revenue per full-time equivalent student
TUITIONFEE_OUT	Out-of-state tuition and fees
UGDS	Enrollment of undergraduate certificate/degree-seeking students (size)
UGDS_API	Total share of enrollment of undergraduate degree-seeking students who are Asian/Pacific Islander
UGDS_ASIAN	Total share of enrollment of undergraduate degree-seeking students who are Asian
UGDS_BLACK	Total share of enrollment of undergraduate degree-seeking students who are Black
UGDS_MEN	Total share of enrollment of undergraduate degree-seeking students who are men
UGDS_WOMEN	Total share of enrollment of undergraduate degree-seeking students who are women

Contributions

To clean the data, a clean_data.py script stored all cleaning functions. All group members contributed to this file. These cleaning functions were organized and captured by a single run_All() function that was used when importing a data set. All group members engaged in the various stages of feature engineering; consulting one another on potential ideas before any implementation.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary

For general code management and version control, we established a github repo https://github.com/wiegan204/College_Project. With Jesse guiding the group, we all learned how to use this software effectively and responsibly.

We divided up the modeling and evaluation responsibilities. Modular and scalable code was written by Khasi and Will in our script modeling_2.py, and we continued to use these throughout our project. Jesse implemented the feature selection methods and models, and Will performed the model tuning for our best-performing model.

Every group member contributed to every section of the writeup, though Data Understanding and Preparation were chiefly organized and implemented by Khasi, Deployment by Jesse, and the Modeling and Evaluation section by Will.

[1]: *Using Federal Data...*2015

[2]: Data Dictionary