Is College Still Worth It?

Predicting Post-College Earnings

Aidan Claffey, Dan Turkel, Kevin Wilson
New York University Center for Data Science

DS-GA 1001 Term Project - December 2019

Abstract

Students facing the college application process often do not know where to start. Those seeking high earnings after college will often use the information on the college admissions website like "median/mean salary after graduation" to help decide which schools will pay off in the long run. However, this data is often already outdated, since applying students want to know what their earnings will be *when they graduate*, not what current graduates' earnings are.

To help students make a more informed decision about how a college will prepare them for a career, we use the U.S. Department of Education's College Scorecard data and several regression methods including Linear Regression, Random Forest, and Gradient Boosted Trees, to (1) better predict the expected earnings of graduates from each college in the dataset and (2) identify the characteristics of a college that are most predictive in determining post-college earnings.

1 Business Understanding

The circumstances surrounding the decision to attend college (and if so, where to apply and enroll) have changed for the next generation of college students. For decades, a bachelor's degree brought with it a more certain financial future, enabled primarily by more lucrative career opportunities. While always an investment, the financial benefits and job opportunities seen by college graduates meant that a bachelor's degree was nearly always worth it.

Today, higher education costs have ballooned, requiring many students to borrow massive sums to attend college. Furthermore, graduates are faced with less certain and less lucrative earnings prospects, driven by slower nominal wage growth, the rise of automation, and accelerating urbanization. [1] [2] A survey by the APM Research Lab [3] found that roughly 70% of American college graduates say the degree is worth the cost, while only 58% of Americans at large think so.

Average in-state tuition and fees for a four-year public college for the 2019-2020 academic year are 2.97 times higher than for the 1989-1990 year, with rates rising at over 2% per year for each of the last 10 years. Meanwhile, median family income in the United States, after adjusting for inflation, is just 19% higher than in 1989. [4] Today, student loan debt in the United States has surged to over $1.5 trillion; one in seven recent graduates have a debt load higher

than their first-year earnings, and one in 50 owe more than twice their first-year earnings. [5]

Students facing the college application process today face a deluge of information that can be difficult to comprehend and use to make a decision as monumental as where to attend college. Today, more than ever, the characteristics of each institution and field of study have a significant impact to earnings potential and whether the tremendous burden of financing such an education is "worth it." Most institutions (in admissions materials, campus tours, and elsewhere) tout career placement and earnings statistics for their graduates, but such data are often outdated, opaque with regards to their sources (e.g. self-reported or survey on opt-in surveys) and difficult to compare.

While the decision of whether and where to attend college should be made based on more than just the presumed financial benefit, the dramatic rise in cost of higher-education makes it invaluable for incoming students to better understand their future prospects.

1.1 Prior Work

Much research has been done into the question of whether college is an economically wise decision. A large-scale study conducted by Georgetown University [6] found that associate's degree and certificate programs lead to the highest short-term returns, and four-year degrees from private non-profit institutions lead to the highest long-term returns. The study found that, even accounting for the higher cost of private college, graduates from such schools generally earn greater lifetime net earnings than public college graduates.

A study [7] of British university graduates using a dataset similar to the College Scorecard sought to examine which traits of a *student* led to the highest earnings, while our analysis examines data on traits of the institutions themselves.

A Stanford University student project [8], discovered by the authors after completing our analysis and modeling, came the closest to our particular task, using the College Scorecard dataset to predict post-college income as a supervised learning problem. The students focused in particular on "32 features provided by the US Department of Treasury, including gender, age, ethnic, and income demographics of students."

2 Data Understanding

The United States Department of Education releases annually its "College Scorecard" dataset [9], containing institution-level data about the 7,112 colleges and universities in the United States (in the most recent release). The data is obtained from various federal agencies, U.S. financial aid application data, and tax information. The most recent dataset, describing the 2017-2018 academic year, was released on November 20, 2019.

The dataset is vast, with 1,978 fields describing institutional characteristics (location, religious affiliation, Title IV eligibility), academic profile (degrees and fields of study offered), admissions statistics (standardized test scores, admission rates, completion and retention rates), student demographic data (race and gender, full-time/part-time, age), and more. Relevant to potential extensions of our problem are several fields decomposing schools' tuition and fee structure, as well as details on financial aid awards, net cost to students, and ultimate cumulative debt incurred by students. The technical documentation [10] and data dictionary [11] that accompanied the data were essential in comprehending some of the more complex features.

The data also include several fields measuring mean and median earnings for various cohorts of students, with such data available for Title IV eligible students (see *Data Caveats* section) disaggregated based on number of years since first enrolling in college. "Earnings" was defined as the sum of wages, deferred compensation, and self-employment income reported on federal income tax returns. [10] Data are available only for students not currently enrolled in school (i.e. graduate students are excluded) and for each year starting 6 years after a student first enrolls in school up to 10 years after a student first enrolls in school.

One critical aspect of the data as it is distributed is that the post-enrollment income is not associated with students who enrolled in the academic year in question. [11] This is to say: in the 2014-2015 dataset, the 6-year post-enrollment data corresponds to students who enrolled six years *prior* to 2014-2015. In particular, it corresponds to a pooled cohort of students who enrolled in the 2007-2008 *and* 2008-2009 academic years. In *Adjusting the Target Data*, we discuss how we adjusted the data to account for this.

2.1 Data Caveats

The College Scorecard data contains several caveats that must be considered. The technical documentation [10] notes "two notable limitations" on earnings figures (the focus of our task) in particular.

The first is that the earnings are aggregated across all programs at an institution. The documentation notes that earnings variation may be greater between programs at an institution than it is from one institution to another. While program-level data is slated to become part of the College Scorecard, it is not yet available (see *Conclusion and Future Work*). This constitutes a large issue as schools have different distributions of student majors, but students consulting the data are often considering the same course of study in all schools. For example, MIT has many more high-earning STEM majors than Vassar does, which inflates its earnings, but this is not relevant to a student considering attending either school for philosophy. It is therefore advisable that students also consider the data that College Scorecard provides on proportion of students enrolled in certain majors when consulting the model results.

Second, earnings data is only collected from Title IV students (whose tax filings are cross-referenced with their financial aid applications) and excludes students still enrolled in university. Consequently, the earnings data are biased towards the outcomes of students who received financial aid, which might bring the earnings data downwards as students who do not apply for financial aid often come from wealthy families and parental wealth has been shown to be a strong predictor of future wealth [12].

Additionally, 6 years after enrolling in an undergraduate program, many students are enrolled in graduate programs and some are still enrolled in undergraduate programs. Depending on the balance between these two groups at a given institution, the exclusion of still-enrolled students from the earnings figures could further bias the data.

Lastly, College Scorecard notes some general limitations of their data. Of particular relevance are the inconsistent cohort definitions across metrics (see *Adjusting the Target Data*). The metrics in the dataset come from a variety of raw sources, many of which were not originally intended to interoperate, and so there is some inconsistency in how the student groups being measured in each metric align.

3 Data Preparation

3.1 Ingestion

Prior to ingesting the data itself, we imported the data dictionary [11] into a Python dataframe to have as a quick reference for column definitions. We also used this dataframe to create a dictionary of categorical feature mappings, since these features were encoded as integers in the raw data.

To import the data itself, we looped through the CSV files of each academic year, imported the CSVs (manually specifying that `\NU"` and `\PrivacySuppressed"` should be treated as nulls), added a column for the year of the data from that CSV, and added a unique ID for each row by concatenating the institution ID with the year.

3.2 Subsetting the Data

The College Scorecard dataset includes data from the 1996-1997 school year to the 2017-2018 school year. We chose to use the 2002-2003 through 2017-2018 data, as the schema changed progressively over time, with some larger changes occurring in the 2000-2001 year. (The 2001-2002 earnings were pooled with 2000-2001, so we chose to exclude them as well.)

The earnings data has not yet been compiled for students who entered college post-2009, so the 2002-2003 through 2008-2009 academic years served as our training and validation data. (Earnings data from the data up through 2014-2015 were used in this dataset because of the 6-year offset between the features and the target variable. This is explained in *Adjusting the Target Variable*.) A subset of those years was also excluded and used for hold-out testing (see *Modeling & Evaluation*), the features for the post-2009 data serve as unlabeled data that our model's users can consult for predictions on future earnings.

We applied some filters to the data as well. Each row of the raw data constituted an institution/academic year pairing, and we only included institution-years where the school had at least 50 undergraduate degree-seeking students, at least 250 Title IV students, was currently operating, were not graduate-only, and offered at least an associate's degree.

3.3 Initial Feature Selection

The data dictionary for the latest version of the College Scorecard dataset features nearly 2,000 columns. Our coarsest form of feature selection involved manually sorting through the data dictionary and selecting variables that seemed reasonably correlated with future income while excluding overly redundant features (e.g. the same statistics cut different ways) and certain demographic columns that seemed out of step with our philosophy of evaluating the institutions rather than the individual students.

For instance, the feature `UG_NRA` encodes the percentage of non-resident aliens in the student body; we felt an individual's citizenship status would be much more predictive of income than the share of a college with a given citizenship status, and since individual demographics are out of the scope of this project, we excluded these types of features.

We also excluded other earnings data (e.g. earnings 10 years post-entry), and information irrelevant to our scope of work, like whether or not a college offered certain certificates.

We ultimately selected 110 columns for consideration out of the original 1977.

3.4 Missing Data and Backfills

There were several cases in the raw data where a value of interest was split across several mutu-

ally exclusive columns. For example, `COSTT4_A` and `COSTT4_P` both encode the average cost for a year at an institution, but split by whether the institution uses an academic calendar to define a year, or a "program year." Other columns were split between public school and private school variants of the same feature. In these cases, we created a new column that combined the source columns so that no institution had a missing value.

We also examined the percentage of rows that had null values for the columns we had chosen. For six numerical features with less than 37% of the data missing, we filled the nulls with the average value (across all years). These features were: proportion of faculty that is full-time, instructional expenditures per full-time equivalent student, average faculty salary, median debt for students who have completed the program, completion rate for first-time, full-time students in 150% of expected completion time, and admission rate. For all of these columns, we created a new boolean column indicating that the value had been imputed (`{col}_ISNA`).

We dropped an additional 24 columns which had null rates greater than 50%. These were primarily percentiles of standardized test performance but also included completion rate, percentage of undergraduates aged 25 and up, and percentage of undergraduates who received a Pell Grant.

For five features that had been included in later years and not others, we backfilled them so that they could be included in prior years. These features were: locale (e.g. small, medium, large), basic Carnegie Classification (e.g. Special Focus Two-Year: Technical Professions, Doctoral/Professional Universities), undergraduate profile Carnegie Classification (a description of length of program, proportion full-time, and selectivity), size and setting Carnegie Classification (e.g. Two-year, small or Four-year, small, primarily nonresidential), and whether the university is currently operating.

For these backfilled features, we felt confident that they were unlikely to change over the years in the dataset.

We also backfilled the number of Title IV students. While these values *are* likely to change throughout the years in our data, we only used those columns to filter out schools with fewer than 250 Title IV students and then dropped the column afterwards, so we felt confident that backfilling from the latest year would not create an issue.

Lastly, we dropped any rows in which there were no earnings data.

3.5 Data Cleaning and Final Touches

In order to differentiate schools that offered in-state tuition, we created a boolean column to indicate when a school has different values for in- and out-of-state tuition. (These two types of tuition were already features in the data.)

Because categorical variables were encoded as integers, we used the lookup dictionary we had built to replace these integers with their actual labels. This led to more useful column names when we applied one-hot encoding to the categorical variables (dropping the first value for each variable to avoid multicollinearity).

Throughout the cleaning process, a few boolean columns with low support had been reduced to all-zero, so we removed these zero-variance columns before the final data export.

Our final labeled data set was 11,201 observations with 200 features each.

3.6 Adjusting the Target Data

Various adjustments were made to individual features as well as to the target variable to account for the time period from which the data were collected. For example, to account for the effect of economic inflation on earnings, we adjusted reported values to 2018 dollars, using as a benchmark the widely-accepted U.S. Bureau of Economic Analysis' Consumer Price Index. For example, the earnings data for 2011 as reported in the dataset had been inflation-adjusted to 2014; our computation involved further adjusting this data to 2018 dollars by multiplying by a factor of 1.06071.

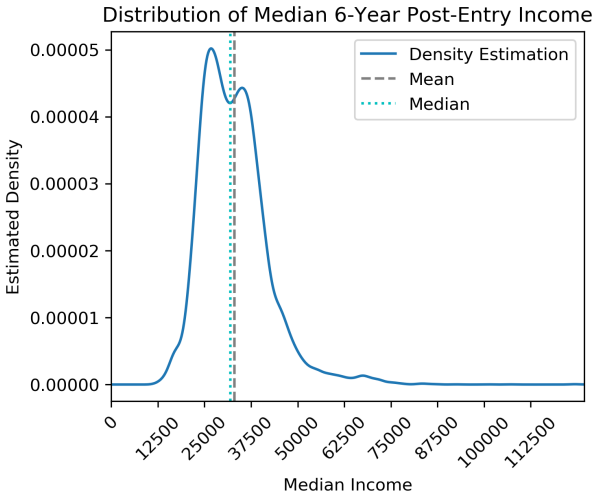
On top of that, we had to reassign the earnings data to the years corresponding to the students who made those earnings. In the raw data for the 2014-2015 academic year, the 6 years post-entry earnings column corresponds to students who had entered the institutions 6 years *prior*. In fact, the figure is a blended average of the students who entered the institution in 2007-2008 and in 2008-2009.

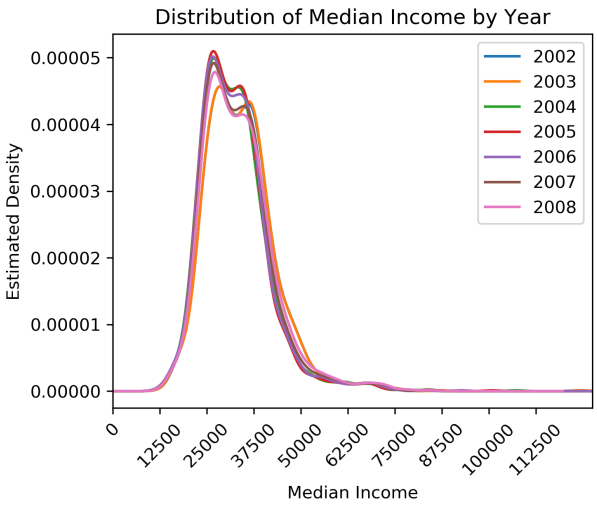
We used the number of students counted in the earnings calculation to create new weighted averages corresponding to the correct cohort year. So our earnings for the matriculating class of 2007, for instance, was a weighted average of the earnings recorded in in 2013 and in 2014. This meant that the target variable for each year necessarily contained some temporal leakage from other years at the same institution, which we were careful to prevent from affecting our model, as described in Section 4.

3.7 Exploratory Analysis

The target variable has a mean of $33,060, with a minimum of $12,508 and a maximum of $124,315. The median is close to the mean, at $31,962.

The schools with the highest median earnings for students matriculating in 2008 contains a number of usual suspects: Duke, MIT, Stanford, Harvard, and so on. However, there are also some more surprising entries: Massachusetts College of Pharmacy and Health Sciences (MCPHS) scores in the top handful of earners every year, likely bolstered by the high job security and high-earning post-graduate professions that a pharmaceutical/medical school offers. The Colorado School of Mines, Stevens Institute of Technology, WPI, RPI, and Georgia Institute of Technology all produce a high number of engineers and technical professionals who are able to quickly obtain higher-than-average salaries.

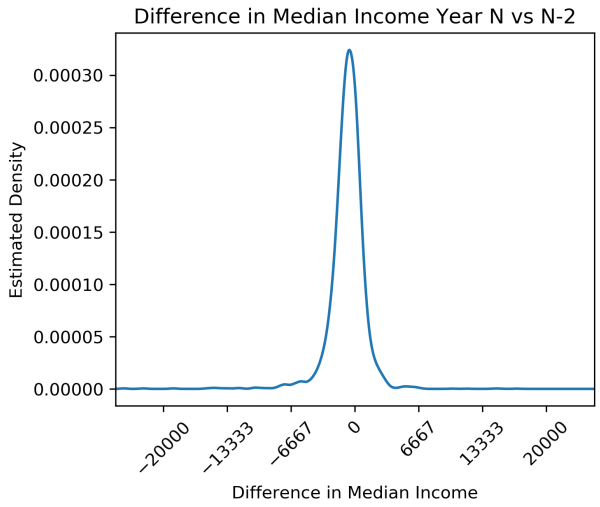




We were surprised to find that the income distri-

butions did not change very much from year to year once adjusted for inflation.

As it turns out, not only is the distribution fairly steady, but we would later discover that individual institutions tended to have very similar earnings outcomes over time. Comparing an institution's median earnings to those of the same institution two years earlier (skipping a year to avoid leakage because of the blended cohorts used to create the income figures, see *Adjusting the Target Data*), we found that the average absolute deviation in income was only $1630. In 97% of cases, the difference was less than $5000 dollars.



This fact contributes to the strong performance of our baseline model, and adds a challenge to the task of making predictions that are better than just looking at last year's data.

4 Modeling & Evaluation

To evaluate model performance, we chose the following regression evaluation metrics: R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These are the most common and most interpretable regression metrics, and since we are building our models to be both accurate and interpretable, these metrics made the most sense. In particular, RMSE and MAE give us error in terms of dollar units, whereas MSE would give dollar-squared units of error.

4.1 Time-Series Cross Validation and Evaluation

Because the College Scorecard data contains entries for the same colleges over multiple years, we had to come up with a way to make sure predictions for past data were not using future data in the calculation (since that would not be replicable in a deployed scenario).

To account for this, we used a method similar to scikit-learn's `TimeSeriesSplit` (discussed in a Towards Data Science blog post [13]). Since our data covers multiple years, we used 6 folds for cross validation, where the each successive fold's training set contained one more year of data and each test set contained the following year's data. For example, the 2nd fold has years 2002 and 2003 in its training data and year 2004 as its test set, while the 3rd fold has years 2002, 2003, and 2004 in its training set and year 2005 in its test set. These folds are used as cross-validation to tune the hyperparameters of the model.

In the case of the Linear Regression model, we skipped the fold with only one year of training data, since the model is unable to learn the coefficient for the year in that case.

4.2 The Baseline Model

Since our dataset has multiple years of data for the same college, a reasonable baseline "model" is predicting the current year's income as the income 2 years back of the same college multiplied by some coefficient α, as shown here ($i_{c,t}$ is income from col-

lege c at year t):

$$i_{c,t} = \alpha i_{c,t-2}$$

We chose to use data from 2 years back instead of 1 year back because the income from 1 year back has direct information about the current year's income (since we used an average between 2 years to create the target variable). We call this model the "Naive Predictor".

The simplest choice for α is $\alpha = 1$ (that is, assuming this year's income will be the same as last year's at the same institution), and a more complex choice for alpha is

$$\alpha = \frac{i_{c,t-2}}{i_{c,t-4}},$$

where the coefficient scales $i_{c,t-2}$ by the change in income from the year $t-4$ to year $t-2$. (Again, we use increments of two years to avoid data leakage from blended cohorts.)

The benefits of the Naive Predictor are that it requires very little computational power, since each prediction only relies on one or two data points, and is extremely easy to interpret as it is premised on the hypothesis that an institution's earnings are stable on a short time-scale. Additionally, it is most likely how students currently use the College Scorecard data to judge colleges based on future potential earnings.

However, the Naive Predictor does not reveal which features of the data affect earnings potential, nor does take into account features of a college other than its name, so this method is likely to perform poorly on schools where large changes are implemented that affect earnings outcomes. Furthermore, if the College Scorecard adds a college to its dataset today and does not have previous years' data, the Naive Predictor will have no data to make a prediction.

To address these issues, we will implement regression models such as Linear Regression, Random Forest, and Gradient Boosted Trees, which will be discussed in later sections.

The Department of Education can be slow to collect post-entry income data (non-earnings income, which is the input to our model, is collected and made available much more quickly), and income from 6 years after entering a school is always delayed by definition. The final outcome of our model is to provide an accurate estimate of median income 6 years post-entry at a college using the college's *current* characteristics. Our end-users would be using our model's predictions for institutions where no recent earnings data has been published or collected, but the data which our model needs to *predict* those earnings is already available.

4.2.1 Naive Predictor Performance

The Naive Predictor performs very well based on the following metrics:

	$\alpha = 1$	$\alpha = \frac{i_{c,t-2}}{i_{c,t-4}}$
R^2	0.9314	0.8838
RMSE	2357	3096
MAE	1630	2274
Missing Values	3184	6388

The missing values indicate how many rows the Naive Predictor could not predict income for because it did not have enough data to look backwards. "Training" the model on older data would help fill in some of these missing values, but would not account for the addition of any new schools over time.

4.3 Linear Regression

First, we wanted to see if there was a linear relationship between the college's features and our target.

Linear regression (LR) is the most accessible model for that task, so we ran one using both the normal data set and a scaled version of the data set, with the following results:

	Unscaled		**Scaled**	
	Training	Test	Training	Test
R^2	0.8443	0.8235	0.8433	0.8209
RMSE	3637	3799	3638	3802
MAE	2582	2807	2588	2846

Clearly, neither the normal nor the scaled LR model performed as well as our naive model in any of the metrics, although it did provide feature coefficients that we use later to determine the sign of the Random Forest model's feature importances.

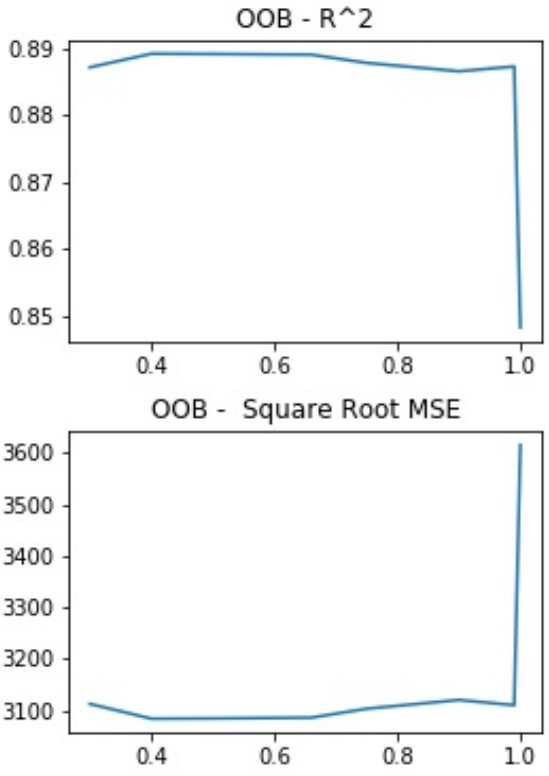
Note that in scaling the model features, we needed to use the `RobustScaler` rather than the `StandardScaler` in scikit-learn. We found that the `StandardScaler` struggled with the boolean features with low support.

4.4 Random Forest

More complex than linear regression, Random Forest (RF) regression provided more accurate predictions with more interpretable feature scores. The Random Forest has the advantage of being able to learn complex nonlinear relationships between the data and the target.

We tested the `max_features` parameter at different values to see which provided the best results, with the idea that leaving out more features gives a lower bias, and since there are 100 trees, the variance gained by having different features is minimal. In particular, of concern was whether some of the one-hot encoded categorical features which turned into many boolean columns would lead to a selection of features in a given tree that were highly correlated, and we felt that using a smaller percent of features per tree might alleviate this.

The bootstrapped RF gives out-of-bag predictions that we used to choose the best `max_features` hyperparameter value.



Besides when `max_features=1`, the model performed similarly at all other levels of the parameter, with a slight optimum between 0.4 and 0.6. So, we chose `max_features=0.5` to use to evaluate the test set, which had the following results:

	Training	Test
R^2	0.9785	0.9091
RMSE	1356	2726
MAE	809	1961

While the test set did not outperform the Naive Model in any of the metrics, the Random Forest outputs feature importances. Since Random Forest

feature importances are unsigned, we made the assumption that the a feature's Linear Regression coefficient acts in the same direction as in the Random Forest. The top fifteen features are displayed in Appendix B, and the top 5 are listed below (along with their feature importance and sign).

C150_4	0.192037	+
TUITIONFEE_OUT	0.147867	+
AVGFACSAL	0.102896	+
PCIP14	0.058292	+
HIGHDEG_Graduate	0.041751	+

These features represent the following:

- C150_4: Completion rate within 6 years for 4-year programs
- TUITIONFEE_OUT: Out-of-state tuition
- AVGFACSAL: Average faculty salary
- PCIP14: Percent of degrees awarded that are engineering
- HIGHDEG_Graduate: Highest degree at institution is graduate-level

These features are discussed further in *Conclusion and Future Work*.

4.5 Gradient Tree Boosting

While Random Forest is an *averaging* ensemble method wherein each tree is learned on the bootstrap sample of the data, Gradient Boosted Trees (GBT) is a *boosting* method by which base estimators are built sequentially based on the cumulative residuals of the predictions of the previous estimators. In each stage a regression tree is fit on the negative gradient of the given loss function (least-squares).

We implemented this method using the same time series-based training and test splits discussed above, with the best performance attained with the following key hyperparameters: Number of estimators (300), loss function (least squares), learning rate (0.1), max depth (1). Note in particular that the default number of estimators is 100 but a higher number was deemed appropriate given GBT is quite robust to overfitting.

The model performed well, but not as well as our Linear Regression or Random Forest models. The best results achieved using this model are included below.

	Training	Test
R^2	0.8298	0.7977
RMSE	3802	4066
MAE	2725	2979

The top feature importances identified by our GBT were similar to those identified by Random Forest, although of different magnitude. C150_4, TUITIONFEE_OUT, and AVGFACSAL were the top-most important features, with respective feature importances of 0.253, 0.226, and and 0.117. Refer to Appendix B for a list of the top 15 feature importances identified by this model.

Overall, this method produced results that are competitive with, but not better than, the other methods we attempted.

4.6 College Holdout Method

All of the above metrics are an average over each time-series fold. Part of the reason they are lower than the Naive Predictor is that the earlier folds have less data to train on, so those predictions are less accurate and bring down the aggregate R^2, $RMSE$, and MAE. We tried a different method of splitting the data into training and test, which we call the "College Holdout" Method. Instead of splitting by time, we split by college, such that 80% of the colleges are in the training set and 20% are in the test set. Therefore, it is still the case that no prediction

for a college is being made with future data about that college.

Using the College Holdout method, our Random Forest performed better than before, with $R^2 = 0.940$ and RMSE$= 2254$, which are better than the Naive Predictor. The feature importances were similar to those in previous model runs.

5 Deployment

As described above, the model performs quite well as a predictive tool in determining the expected earnings a prospective student might attain after attending a college given a set of characteristics of that college. In order to provide end user usability for this model (that is, to implement it in such a way that prospective students and other interested parties might use it to improve their decision-making), the model must be embedded in a user-friendly, web-based portal on which users can input either specific colleges they are interested in attending, and/or characteristics they are seeking in the process of narrowing their college search, and the model will output the predicted earnings given such characteristics as well as a mechanism for comparison.

Notably, the College Scorecard system does include some degree of search functionality, but with several limitations, including (1) the features available to be selected are quite limited; (2) the tool provides no mechanism to sort or search by post-college earnings; (3) the data are static (reflecting only the most recent dataset) and incorporate no modeling/predictive aspect such that future earnings can be estimated or inferred. Deploying our model effectively would enable students to see which characteristics are most important in predicting future earnings, weigh their own preferences against those characteristics, and estimate earnings at the colleges that fit their specifications.

5.1 Risks, Issues and Ethical Considerations

A discussion of the risks and issues surrounding deployment and implementation of this model is of paramount importance (for we expect that this model could serve to improve one of the most critical decisions of a young person's life). We note that, while the model performed well on the test data, the reality of the political and economic landscape in the United States is uncertain, and given the model is based solely on past results there are no guarantees that future earnings are certain to be attained based on attending a certain school. Acute events beyond the predictive scope of this model could also have an impact; one example of this is the college admissions bribery scandal exposed in March 2019 [14] likely reduced the attractiveness of Yale, Stanford, USC, Wake Forest, and Georgetown, as well as the hirability of student-athlete graduates from these schools (which are otherwise associated with high earnings potential).

Additionally, there are several considerations from a data responsibility and ethics standpoint. We would be remiss not to give mention to the significance of race, gender, and socioeconomic status in the college admissions process, and in this dataset. The raw data included several features that would be problematic to include as determining factors in a prospective student's college selection process. These include, among other features that we eliminated, the set of racial/ethnic classification variables (i.e. `UGDS_BLACK` and `UGDS_ASIAN` which quantify the total share of enrollment of undergraduate students who are Black and Asian, respectively, in a given

year), and the set of gender classification variables (i.e. `UGDS_WOMEN` which quantifies the total share of enrollment of undergraduate students who are women, in a given year).

Feature importances associated with those variables should not be used to encourage or discourage students from applying to or attending a specific school. Furthermore, the relationship between demographic data and the earnings of the graduates is likely affected in large part by long-standing societal biases. Students from groups which are discriminated against will not necessarily gain any earnings advantage from attending a school where the majority of students are from a privileged group.

To be clear, the issue is nuanced and admissions departments differ in their treatment of race, gender, and other demographic characteristics. Diversity of such backgrounds in a student body almost certainly leads to better outcomes. However, given the intention for the model to be used chiefly by students selecting which colleges are "worth" applying to and ultimately enrolling in, it felt most prudent to remove these fields from that calculation.

6 Conclusion and Future Work

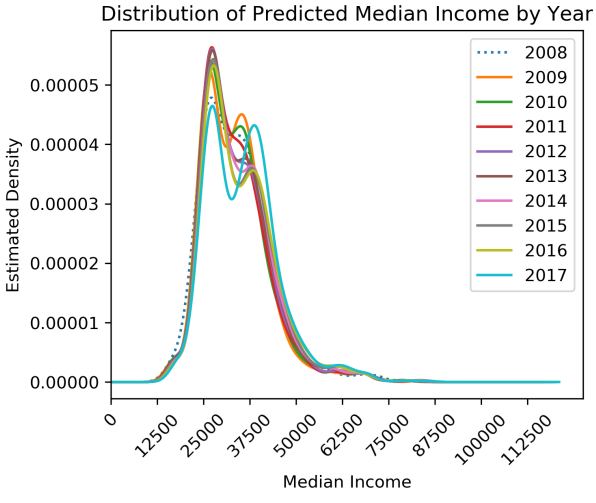
The decision-making involved in the process of applying for and attending college, and the earnings one might hope to attain after doing so, is complicated but incredibly important. By examining this problem data-analytically, and implementing several regression methods including Linear Regression, Random Forest, and Gradient Boosted Trees, we were able to establish, with reasonably strong performance, which characteristics of a college are most predictive in determining post-college earnings, and develop a mechanism for predicting earnings based on the characteristics of a given school.

6.1 Conclusion

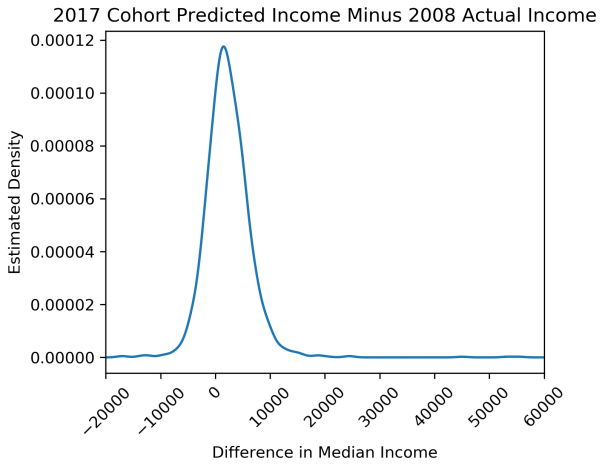
The baseline model showed that our task has a high bar for machine learning to beat the "naive" approach to the problem, but our machine learning methods were able to produce highly competitive results in scenarios where the naive model cannot be applied (new schools, schools with highly varying features over time, and schools in the years before the earnings data becomes available for a given cohort).

We ran our best Random Forest model on the unlabeled data from the cohorts entering college in 2009-2010 onward. The figure below compares the predicted income distributions to the true distribution from 2008-2009.

The predicted distributions are all slightly (but only slightly) to the right of 2008's, which may come as a relief to future college students. The median incomes for these years hovers between $32,000 and $33,000, with a jump to $35,000 in 2017. The jump in the last year may be due to more missing data in the most recent year adding variance to the model predictions (as some of the data likely takes longer to collect).



Comparing the predicted 2017 outcomes to the 2008 true earnings data, we measured a mean predicted increase of $2324 and a median predicted increase of $1986. Some extreme differences appeared to be related to schools with identical names, which threw off our ability to match the correct schools over years.



The features our Random Forest model found most important were a mix of sensible and surprising. It makes sense that completion rate would be important, since the dataset includes incomes from those who do not complete the program. Schools with high rates of dropout would then be measuring the incomes of many students who did not end up earning the degree.

It is also reassuring that the out-of-state tuition is an important feature, since one would hope that the more you pay for the education, the more you get out of it. This is likely somewhat explained by the inclusion of some community colleges which only grant associate's degrees and are very inexpensive, and may also feature lower post-graduate earnings. Similarly reassuring was that higher average faculty salaries led to better post-graduate earnings: the best professors will likely not work at colleges where they won't be paid well, but there also may be a demographic aspect: universities in places with very low cost of living (and thus presumably lower salaries) may not have the prestige of universities in urban or more populous areas, and, while not examined in this paper, prestige seems to make an impression on would-be employers.

The percent of awarded degrees in engineering is also an interesting feature. Anecdotally we hear that STEM degrees lead to higher-paying jobs, and this result seems to confirm that idea. But it is also a sign that our model would benefit from program-specific predictions, since the higher average salaries from universities with many engineers lead to inaccurate earnings predictions for those interested in attending that school for a non-engineering program.

Lastly, the fact that a school offers graduate degrees is an interesting feature to see as having high importance. It likely increases the chance that an institution has research efforts, but is slightly non-intuitive since the earnings of graduate students are not included in the College Scorecard post-graduate earnings numbers. There is a chance that schools which offer bachelor's but not graduate degrees are more likely to be small or community colleges, but it warrants further investigation.

6.2 Future Work

In the future, this problem can be extended to incorporate the cost of tuition and fees to compute a "net cost" of attending college, or alternatively a more long-term-focused "return on investment" based on a projection of long-term future earnings. We did not include these complexities in our analysis, (1) given the cost of attending a given institution varies widely even among the same cohort (with some students paying full boat, others getting a full ride, and many in between financing through a combination of loans

and scholarships) and (2) given the difficulty in predicting, and much wider variance in, long-term earnings. That being said, data on student debt and tuition cost is included in the College Scorecard dataset and is worth consideration for future studies.

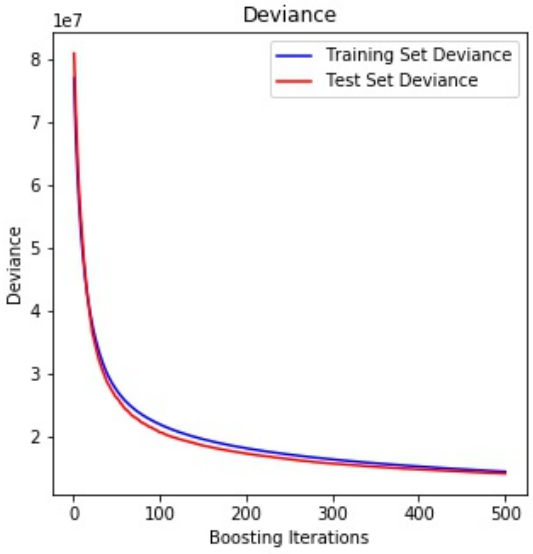
As a target variable, we used earnings six years after beginning school. College Scorecard also keeps data for seven, eight, nine, and ten years after beginning school. A future model could not only make accurate predictions about those time frames as well, but could also see which colleges prepare students best for income growth by using year over year income growth, rather than income, as the target variable. Since many "big-name" schools help students initially get a well paying job, they may not necessarily be preparing students in the long-run. This sort of analysis would help us understand this issue more.

Going forward, the U.S. Department of Education plans to include program-level data which would enable the disaggregation of earnings data across fields of study. Given the fact that a student's field of study and work is highly correlated with earnings magnitude, an analysis of this data in the future would likely add additional value to this decision-making. The reality is that post-college earnings, especially at schools with many programs of study, are high-variance, and anything we can do to further narrow down a particular student's scenario would be helpful to more accurately predict individual outcomes.

Further along those lines, our data makes no attempt to include information on an individual applicant: our model outputs the same income regardless of your race, age, gender, and so on. For as long as these factors remain entwined with earnings, they are relevant to the task of prediction. Additional data would be needed (possibly involving some of the demographic features we excluded in this analysis) to predict earnings at an institution *given traits of an individual in question*.

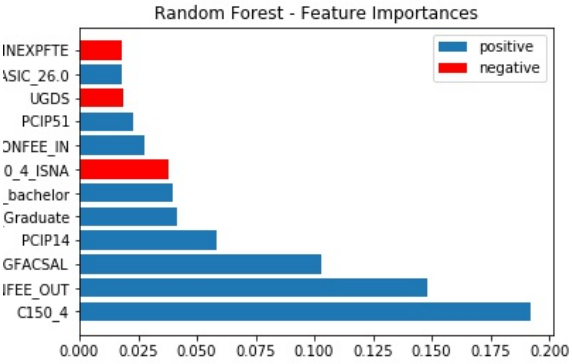
Appendix A Team contributions

Task	Contributors
Data Understanding	All
Data Preparation	Dan
Target Variable Adjustment	Aidan and Kevin
Exploratory Analysis	Dan
Baseline Model	Dan
Linear Regression	Aidan
Random Forest	Aidan
Gradient Boosted Trees	Kevin
Deployment and Risks	Kevin
Conclusion and Future Work	All

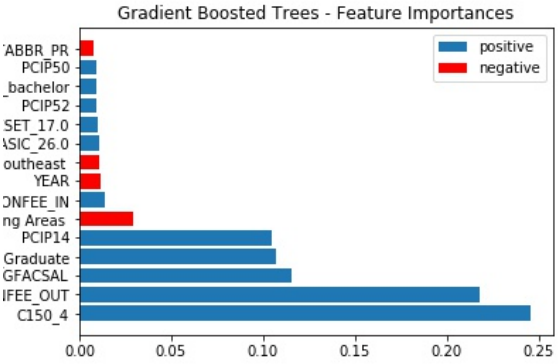


GBT deviance over boosting iterations

Appendix B Figures



Random Forest Feature Importances - Top 15



Gradient Boosted Trees Feature Importances - Top 15

References

[1] Economic Policy Institute. *Nominal Wage Tracker*. Nov. 2019. URL: https://www.epi.org/nominal-wage-tracker/.

[2] PricewaterhouseCoopers LLP. *Five Megatrends And Their Implications for Global Defense & Security*. Nov. 2016. URL: https://www.pwc.com/gx/en/archive/archive-government-public-services/publications/five-megatrends.html.

[3] APM Research Lab. *A First Try at ROI: Ranking 4,500 Colleges*. Mar. 11, 2019. URL: https://www.apmresearchlab.org/freehighered.

[4] College Board. *Growth in published charges*. Nov. 2016. URL: https://research.collegeboard.org/trends/college-pricing/figures-tables/growth-in-published-charges.

[5] Josh Mitchell, Andrea Fuller, and Michelle Hackman. "Which College Graduates Make the Most?" In: *The Wall Street Journal* (Nov. 20, 2019). URL: https://www.wsj.com/articles/which-college-graduates-make-the-most-11574267424.

[6] Georgetown University Center on Education and the Workforce. *A First Try at ROI: Ranking 4,500 Colleges*. Nov. 14, 2019. URL: https://cew.georgetown.edu/cew-reports/collegeroi/.

[7] Chris Belfield et al. *The relative labour market returns to different degrees*. June 2018. URL: https://www.gov.uk/government/publications/undergraduate-degrees-relative-labour-market-returns.

[8] Miranda Strand and Tommy Truong. *Predicting Student Earnings After College*. Tech. rep. Student project. Stanford University, 2015. URL: http://cs229.stanford.edu/proj2015/209_report.pdf.

[9] U.S. Department of Education. *College Scorecard Data*. Nov. 2019. URL: https://collegescorecard.ed.gov/data/.

[10] U.S. Department of Education. *Technical Documentation: College Scorecard Institution-Level Data*. Nov. 2019. URL: https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf.

[11] US Department of Education. *College Scorecard Data Dictionary*. May 2019. URL: https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx.

[12] Ben Casselman and Andrew Flowers. *Rich Kids Stay Rich, Poor Kids Stay Poor*. Feb. 2016. URL: https://fivethirtyeight.com/features/rich-kids-stay-rich-poor-kids-stay-poor/.

[13] Courtney Cochrane. "Time Series Nested Cross-Validation". In: *Medium - Towards Data Science* (May 19, 2018). URL: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9.

[14] U.S. Department of Justice. *Investigations of College Admissions and Testing Bribery Scheme*. Mar. 12, 2019. URL: https://www.justice.gov/usao-ma/investigations-college-admissions-and-testing-bribery-scheme.