Determining the Importance of Institutional & Student Characteristics for Student Future Earnings

Maryam Khalid Shah
McCourt School of Public Policy, Georgetown University

PPOL-565-01: Data Science II

Professor NaLette Brodnax

May 14, 2021

I. EXECUTIVE SUMMARY

Given the importance of future earnings as a factor in college selection and as an indicator of college quality, it is important to understand what institutional and student characteristics are important in determining students' future earnings. Using the College Scorecard data, this project aimed to determine the institutional and student characteristics that are the most important predictors of students' future earnings being in the top 25% across all colleges. For this purpose, the parametric technique used was Logistic LASSO Regression, while the non-parametric technique used was a Decision Tree Classifier. The analysis was conducted for both private and public educational institutions collectively, as well as separately for only private non-profit educational institutions. Average faculty salary and median household income were two of the most important predictors across all models. Factors under the control of educational institutions can essentially be modified in order to increase the probability of their students' earnings being in the top 25%. However, ethical and even legal concerns need to be kept in mind while doing so.

II. INTRODUCTION

For many students, post-graduation earnings, which are a main component of a program's return on investment, are an important consideration while deciding which program to choose. Nearly all colleges report the average earnings of former students on a common scale, allowing for easy comparison. Given the importance of future earnings as a factor in college selection and as an indicator of college quality, it is important to understand what institutional and student characteristics are important in determining students' future earnings. Specifically, what institutional and student characteristics are the most important predictors of students' future earnings being in the top 25% (of earnings across all colleges)? This is essentially both an inference and prediction question since it would provide insights into the current factors that led to earnings being (or not being) in the top 25% as well as help predict whether or not future earnings will be in the top 25% across all colleges given data about institutional and student characteristics.

There have been various studies exploring the relationship between institutional and student characteristics and earnings. Rumberger and Thomas (1993) explored the effect of three variables including both institutional and student characteristics on students' future earnings while James et al. (1989) tried to understand which college characteristics or other aspects of college experience lead to more earnings. Some studies and student projects such as the one by Strand and Truong (2015) have attempted to build models to predict the earnings of college students after graduation. My approach differs from these studies in both the specificity and nature (prediction and inference) of my research question, as well as in the models I employ to answer the research question.

III. DATA

The type of data that would be ideal for analyzing this question is record data, where the dataset would be a collection of records (colleges) which would have a fixed set of attributes, including both student and institutional characteristics. The dataset that was found to best meet the criteria of an ideal dataset was the most recent version of the College Scorecard data (U.S. Department of Education, 2021a). This dataset provides data at the institution-level and heavily draws upon data from the Integrated Postsecondary Education Data System (IPEDS) by the Department of Education's National Center for Education Statistics, the National Student Loan Data System (NSLDS), and the U.S. Department of the Treasury. It includes data on features such as overall admission rate, average SAT and ACT scores, average cost of attendance, proportion of faculty that is full-time, student debt levels, average age of entry, and student earnings among many others that were used to analyze the research question. Data acquisition and preprocessing details are included in the Implementation Appendix.

Table 1: Frequency Tabulations (*N = 1209*)

Control	Total Number
Private Non-profit	756
Public	448
Private For Profit	5

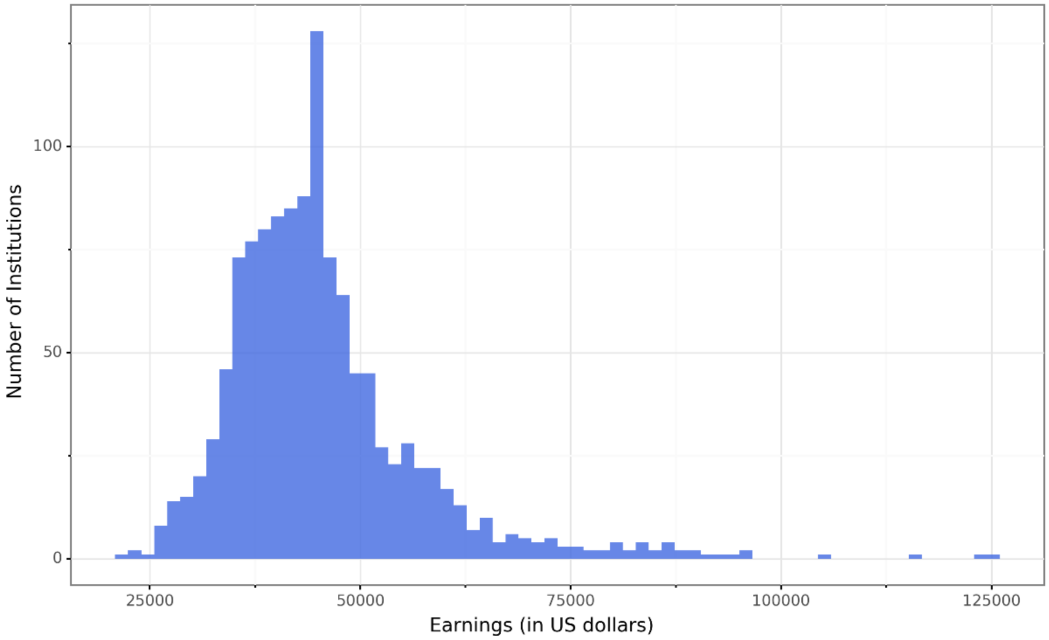
As shown in Table 1 above, the dataset comprises of both public and private educational institutions, with the fewest observations from for-profit private institutions. This low number may be because of the impact of the Gainful Employment (GE) Rule of 2014 – even though GE has now been rescinded, it caused some for-profit colleges to go out of business and others to apply to become nonprofit institutions (Franklin University, n.d.).

Table 2: Descriptive Statistics (*N = 1209*)

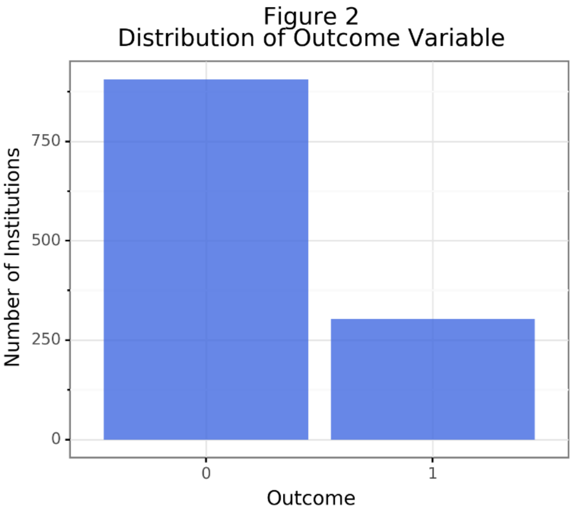
Variable	Mean	Std. Dev	Min	Median	Max
Average SAT Score	1143.75	125.98	785.00	1125.00	1566.00
Median Earnings 10 years after entry	45,474.42	11,651.60	21,100.00	43,600.00	124,700.00
Average Age at entry	21.65	2.32	19.00	21.00	33.00
Average Cost of Attendance	37,339.10	15,731.83	6646.00	36,371.00	75,735.00

Table 2 includes a few of the most important variables, comprising of both student characteristics such as students' age at entry and SAT scores, as well as institutional characteristics like the cost of attendance. The values of all these variable seem to be realistically possible e.g. the maximum SAT score is within the maximum points (1600) a student can achieve. The standard deviations for students' future earnings and the cost of attendance seem to be high, indicating that the observations for these two variables are widely spread. This makes sense since the cost of attendance not only varies between private and public educational institutions but also within each type of institution, and since there is a large difference between minimum and maximum earnings.

Figure 1
Student Earnings(10 years After Entering College)



As can be seen in Figure 1, most institutions' students have future earnings around $45,000 (similar to the average earnings shown in Table 2). This figure shows the distribution of the target variable before it was transformed into a binary variable. The data is skewed towards the right, indicating the presence of clear outliers. Since these outliers do not indicate erroneously entered data and were considered important to the model, they were not dropped from the dataset. Figure 2 below shows the distribution of the target variable after it was transformed into a binary variable by dividing it into two classes, with the top 25% of earnings across colleges coded as 1, and the rest as coded as 0.



As can be seen in Figure 2, more than 750 observations (906 to be precise) were coded as 0, and 303 observations were coded as 1. This resulted in imbalanced classes, and would be talked about in more detail in the sections ahead.

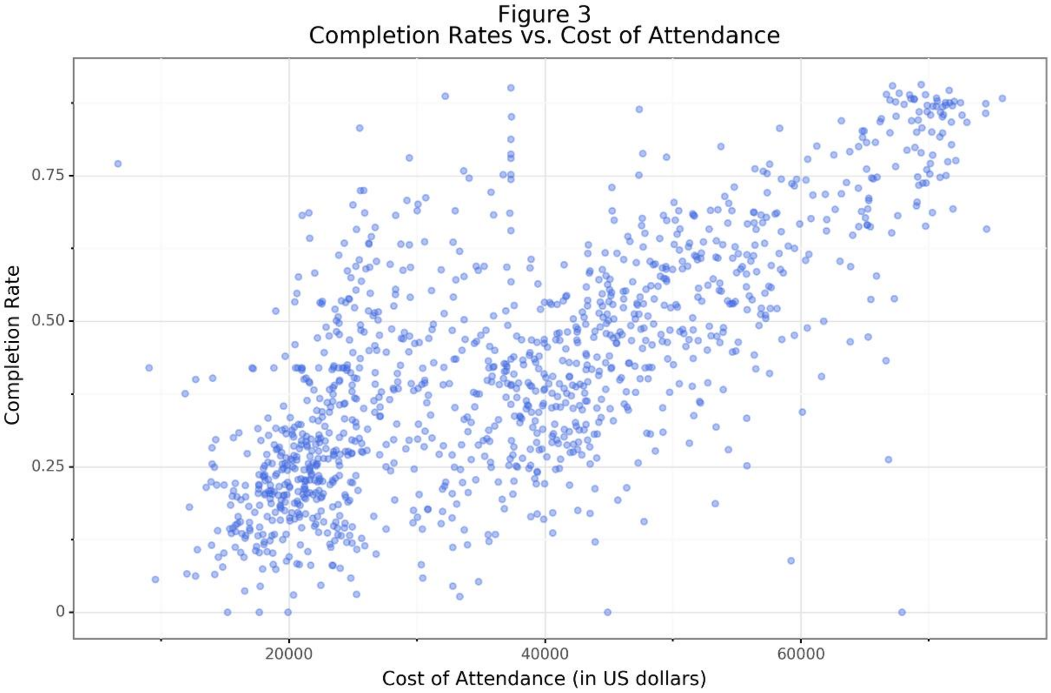


Figure 3 depicts a positive correlation between the average cost of attendance of an institute and its completion rate, indicating that the higher the cost of attendance, the greater the completion rate. A few other pairs of variables were also found to be highly correlated, which meant that there might be difficulty in distinguishing between the variables' individual effects on the target variable. Variance Inflation Factor (VIF) was therefore used to detect multicollinearity, and some variables, including the two plotted in Figure 3, were found to have very high VIF values indicating multicollinearity in the data. Since this could lead to skewed or misleading results in a regression context, it was considered while deciding which techniques to use.

One of the limitations of using the College Scorecard data to answer the research question is that the earnings are estimated for undergraduate Title IV recipients only which means that where the share of each institution's entering class represented by Title IV students is low, results may be less representative of the entire student body (U.S. Department of Education, 2021b). Moreover, the earnings data is provided for all students regardless of whether they completed the program or not so the data cannot be analyzed for only those students who completed the program. This limitation however means that completion rate can be included as a possible predictor of students' future earnings in the machine learning model/s. Moreover, the data does not include information on student characteristics such as academic performance, which has been found to be a significant predictor of students' future earnings (Watts, 2020). However, since SAT scores have been found to be a significant predictor of academic performance (Cornwell et al., 2009), it can be said that academic performance has been partially accounted for in the dataset. These limitations were kept in mind while trying to answer the research question.

IV. METHODOLOGY

In addition to the target variable, selected variables included attributes measuring student characteristics and institutional characteristics such as the average admission rate, average SAT scores, median debt, state the institution is located in, as well as the share of female and first generation students among others. My goal was to both predict future student earnings given institutional and student characteristics, as well as to understand which of these characteristics have the most influence on future student earnings being in the top 25%.

Given that my research question was both prediction and inference, one parametric model that I used with the binary target variable was logistic LASSO regression. Logistic regression makes some tweaks on linear regression by forcing the outcome to fall within the 0 to 1 range, and rearranging it so that it has a linear form, while LASSO is a shrinkage method. It works by adding a penalty term to the log likelihood function and selects from a set of variables in the regression, resulting in a more relevant and interpretable set of predictors (Tibshirani, 1996). In addition to being able to identify the most important predictors, logistic LASSO regression also deals with the multicollinearity in the dataset, by selecting one of the multicollinear predictors (if it is important in determining the outcome) and shrinking the other one to zero. Moreover, outliers in the data have less of an effect on logistic regression as compared to linear regression, since the logistic regression decision boundary only takes the points closer to it into consideration (outliers may still have an effect on the decision boundary however). A limitation of logistic regression is that it does not support imbalanced classification directly. In order to address this and to create equal class representation in the training set, the Synthetic Minority Oversampling technique was used. Moreover, since regularization techniques such as LASSO are not equivariant under scaling of the inputs (Hastie et al., 2001), the variables had to be standardized before running the model.

Logistic LASSO regression has been used to answer various research questions. Pereira et al. (2015) used this technique to predict corporate bankruptcy. Their goal was to improve prediction accuracy, and they believed that LASSO would help achieve this since it deals with multicollinearity and displays “the ideal properties to minimize the numerical instability that may occur due to overfitting” (Pereira et al., 2015).

A non-parametric model that I used given the binary target variable is a decision tree classifier. The objective function of a decision tree is to maximize purity or homogeneity in every node or in as many nodes as possible. Decision tree algorithms are usually greedy, making locally optimal decisions to determine which attribute to use to split the data on at each step in the process. Hunt’s algorithm is one of the simplest algorithms, which uses recursive binary splitting and has two main design elements – the splitting and the stopping criterion. Using a decision tree classifier allowed both prediction and inference since decision trees are relatively easy to interpret as compared to other techniques like Random Forest, as well as allow extraction of feature importance. As there are outliers in some variables such as the target feature, an additional reason of using a decision tree classifier was that it would be less impacted by these outliers.

One main limitation of using a decision tree classifier is that decision trees are prone to overfitting. However, potential overfitting was addressed by setting an optimal maximum tree depth based on a grid search and validation curves. Another limitation is that interaction variables can be problematic in decision trees. This is because a decision tree makes a set of locally optimal decisions, and the interaction variables can be used to split the data in different branches of the tree while the model has no way of accounting for this, or of representing the interaction.

Mashat et al. (2012) used the ID3 algorithm to construct a decision tree for the admission system of King Abdulaziz University (KAU) in Saudi Arabia. Their objective was to build an efficient classification model with high recall under moderate precision in order to improve the efficiency of the admission process (Mashat et al., 2013). Since it was important for the KAU Admissions Office to understand the "overall process", ease of interpretability was stressed as one of the main advantages and reasons of using a decision tree classifier.

V. FINDINGS

This section would be divided into two main parts: an overall analysis for all educational institutions, and a separate analysis for the most prevalent educational institutions – private non-profit, as shown in Table 1. Each of these parts would comprise of two main sections, covering the parametric and non-parametric technique used, namely Logistic LASSO regression and Decision Tree Classifier.

Overall Analysis: All Educational Institutions

The overall analysis was done for all private non-profit, public, and private for-profit educational institutions.

Logistic LASSO Regression

The dataset was split into a training and test dataset in a 3:1 ratio respectively with a seed of 123, after which each dataset containing the predictors (training and test) was standardized separately. Three methods were then used to address imbalanced classes in order to identify the best one. These methods included oversampling the minority class, undersampling the majority class, and generating synthetic samples using the Synthetic Minority Oversampling Technique. The methods were compared on evaluation metrics such as accuracy, precision, recall, AUC (area under curve), and F1 score. The Synthetic Minority Oversampling technique performed the best since it had the highest accuracy, precision, recall and F1 scores among the three methods. After oversampling the minority, an optimal alpha or penalty of 0.3 was identified through a grid search. A logistic regression model from the sklearn package (Pedregosa et al. 2011) was then run with the parameter 'penalty' set to L1 for LASSO, and alpha or 'C' set to 0.3. The 'solver' parameter was set to 'liblinear' as it is a good choice for small datasets, and can also handle L1 penalty (Pedregosa et al. 2011). After the model was run, 14 coefficients out of 68 were zeroed out.

Table 3: Logistic LASSO Regression Performance

Evaluation Metric	Value
Accuracy	82.2%
Precision	60.4%
Recall	81.3%
F1 Score	69.3%
AUC	90.3%

Table 3 shows a few evaluation metrics that were used to assess the performance of the model. The accuracy metric shows that the model correctly labeled 82.2% observations out of all the observations, while the high AUC (Area Under Curve) value of 90.3% shows that the model is fairly good at distinguishing between the two classes.

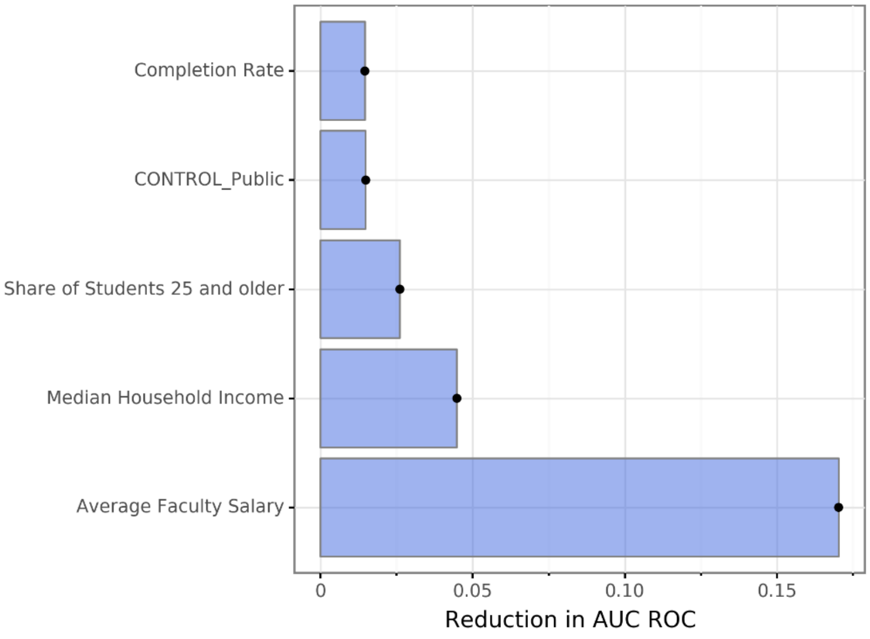
Keeping in mind the research question, both true positives and true negatives are important while the cost of false positives and false negatives is similar. Therefore, accuracy is the preferred evaluation metric.

Table 4: Logistic LASSO Regression Confusion Matrix

	Negative (predicted)	Positive (predicted)
Negative (actual)	188	40
Positive (actual)	14	61

The confusion matrix in Table 4 shows that the most common type of error for the logistic LASSO regression model is of false positives. This is when the actual value is negative but the model predicts a positive value e.g. future earnings of an educational institution's students are not in the top quartile, but the model predicts them to be in the top 25%.

Figure 4
Feature Importance



The features that were the most important i.e. that had the greatest effect on the model's predictions are shown in Figure 4 above. The most important feature was average faculty salary, followed by median household income and the percentage of students who are 25 and older. In order to understand how these variables actually affect predictions, partial dependency plots using the sklearn package were created for the three most important predictors. These plots, as well as the positive values of the coefficients on these variables, showed that as the values of the predictors increase, the probability of an educational institution's students' earnings being in the top 25% increases as well. Two of the three most important features are actually institutional characteristics which educational institutions can review in order to improve their students' future earnings e.g. educational institutions could increase the salaries paid to their faculty.

Decision Tree Classifier

The dataset was split into a training and test dataset in a 3:1 ratio respectively with a seed of 123, after which a grid search was run in order to find the optimal maximum depth from a range of 1 to 10. This resulted in an optimal maximum depth of 3, which was also verified through validation curves. The decision tree classifier from the sklearn package (Pedregosa et al. 2011) was then trained using the training dataset, while its performance was evaluated using the test dataset.

Table 5: Decision Tree Classifier Performance

Evaluation Metric	Value
Accuracy	83.8%
Precision	72.4%
Recall	56.0%
F1 Score	63.2%
AUC	82.6%

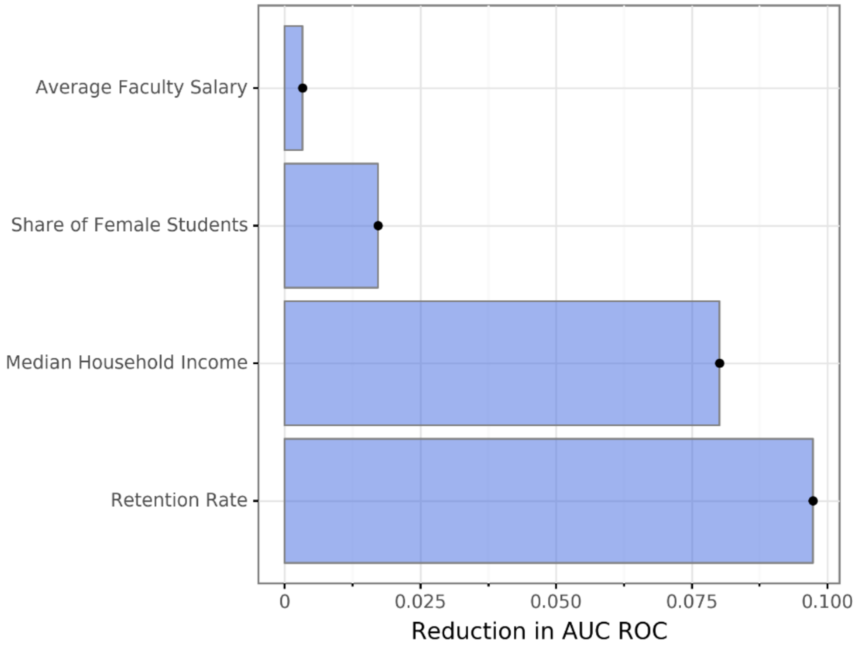
Table 5 shows a few evaluation metrics that were used to assess the performance of the model. The accuracy metric shows that the decision tree classifier correctly labeled 83.8% observations out of all the observations, while the relatively low value of recall shows that out of all educational institutions with student earnings in the top 25%, only 56% were correctly predicted.

Table 6: Decision Tree Classifier Confusion Matrix

	Negative (predicted)	Positive (predicted)
Negative (actual)	212	16
Positive (actual)	33	42

The confusion matrix for the decision tree classifier in Table 6 shows that the most common error was of false negatives i.e. when the actual value is positive but the model predicts a negative value. This was also evident from the low value of recall. The confusion matrix also shows that the model has high specificity i.e. a high proportion (93%) of negative values are correctly identified.

Figure 5
Feature Importance



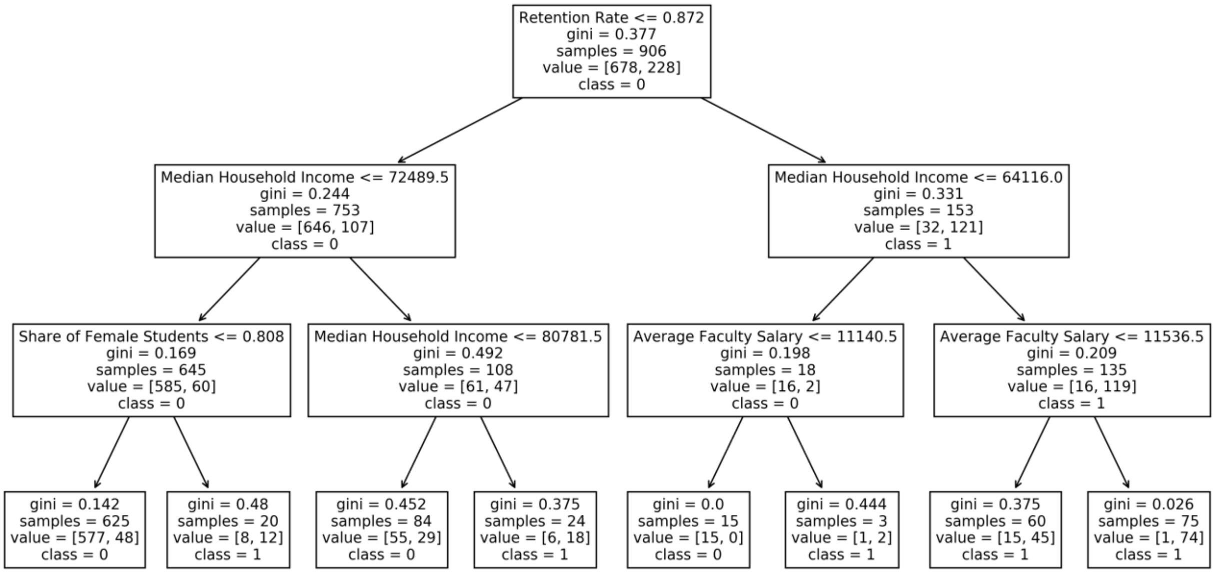
The most (and only) important features in the decision tree classifier were student retention rates, the median household income, the percentage of female students and the average faculty salary. While the average faculty salary was the most important feature in the logistic

LASSO regression model, the reason why it was not the most important in the decision tree classifier is probably because of the different underlying objective functions of the two techniques. Partial dependency plots showed that higher retention rates as well as higher median household incomes increased the probability of an educational institution's students' earnings being in the top 25%.

This means that students for whom high post-graduation earnings matter, can also look at retention rates while considering which college to apply to or join. Colleges on the other hand, can try to improve their student retention rates in order to increase the probability of their students' earnings being in the top 25%, which in turn would attract more students to apply to these educational institutions (given that high post-graduation earnings is one of the factors students are considering while deciding which colleges to apply to).

We can also derive further insights from a visualization of the decision tree, as shown in Figure 6 below. It shows that if the retention rate is less than or equal to 87%, median household income is less than or equal to $72,490, and the share of female students is greater than 81%, student earnings are predicted to be in the top quartile. On the other hand, if retention rate is greater than 87%, the median household income is less than or equal to $64,116, and the average faculty salary is greater than $11,141, student earnings are predicted to be in the top quartile.

Figure 6



Further Analysis: Private Non-profit Institutions

In this part, analysis was done for only private non-profit educational institutions.

Logistic LASSO Regression

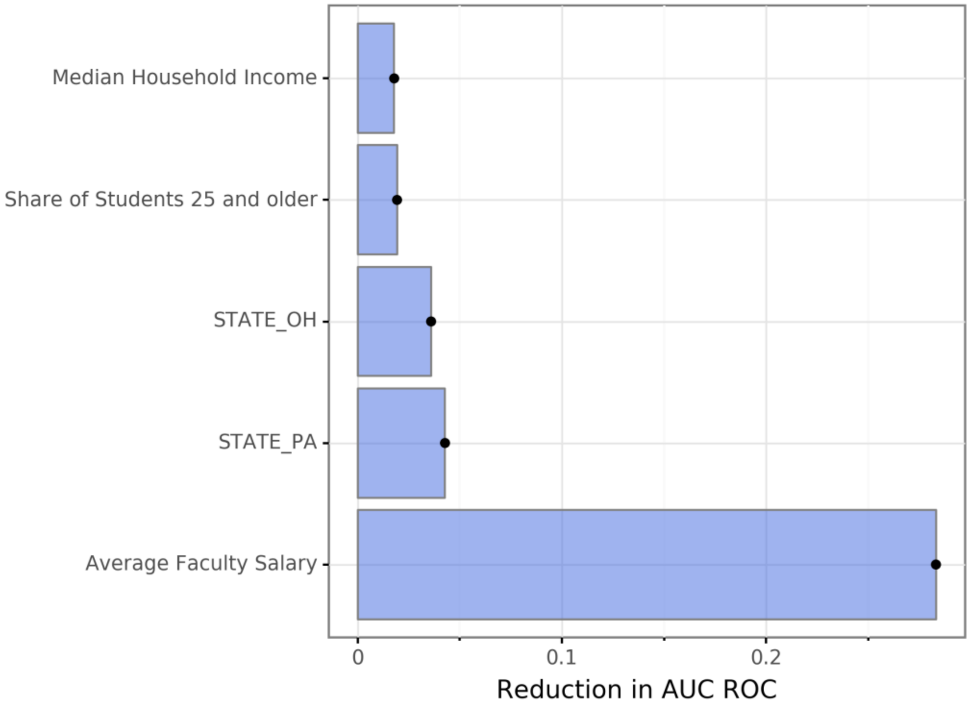
The dataset was split into a training and test dataset in a 3:1 ratio respectively with a seed of 123, after which each dataset containing the predictors (training and test) was standardized separately. As before, three methods were then used to address imbalanced classes in order to identify the best one. The Synthetic Minority Oversampling technique performed the best since it had the highest accuracy among the three methods. After using SMOTE, an optimal alpha or penalty of approximately 0.3 was identified through a grid search. A logistic regression model from the sklearn package (Pedregosa et al. 2011) was then run with the parameter 'penalty' set to L1 for LASSO, and alpha or 'C' set to 0.3. After the model was run, 25 coefficients out of 63 were zoroed out.

Table 7: Logistic LASSO Regression Performance

Evaluation Metric	Value
Accuracy	84.5%
Precision	67.9%
Recall	77.6%
F1 Score	72.4%
AUC	91.2%

Table 7 shows a few evaluation metrics that were used to assess the performance of the model. The accuracy metric shows that the decision tree classifier correctly labeled 84.5% observations out of all the observations, while the relatively low value of precision shows that out of all educational institutions that the model predicted to have student earnings in the top 25%, 67.9% actually have student earnings in top quartile. Once again, since the cost of false positives and negatives is similar in the given context, accuracy is the preferred evaluation metric.

Figure 7
Feature Importance



The most important features for this model (as shown in Figure 7) were the average faculty salary, followed by Pennsylvania state, Ohio state, the percentage of undergraduates aged 25 and above and the median household income. All these features had positive coefficients, which means that as they increase, the probability of the modeled outcome increases (Miller, 2019). Only two of these features seem to be under an educational institution's control – the average faculty salary and the share of students 25 and older, while the location of the educational institution (either in Pennsylvania or Ohio) is realistically not under an established institution's control.

Decision Tree Classifier

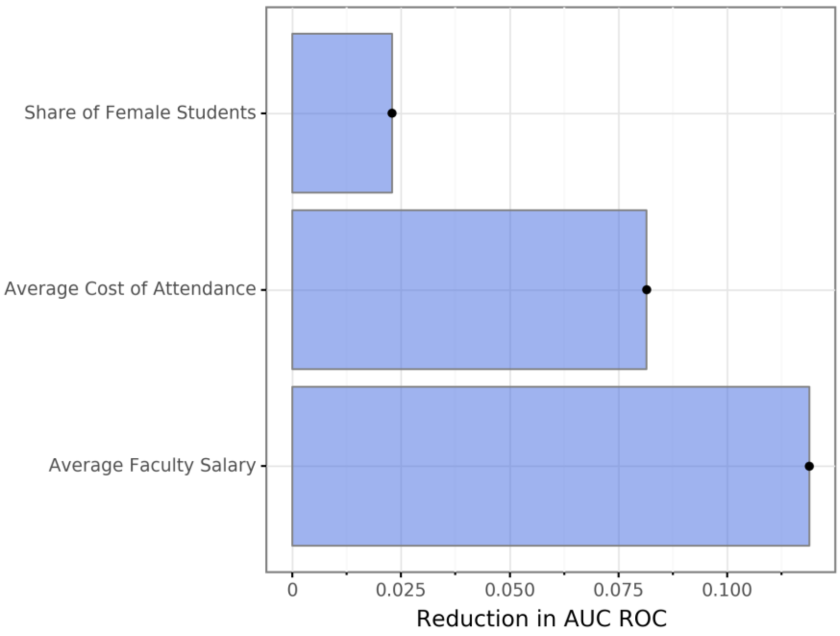
The dataset was split into a training and test dataset in a 3:1 ratio respectively with a seed of 123. Through a grid search, an optimal maximum depth of 3 was obtained. The decision tree classifier from the sklearn package (Pedregosa et al. 2011) was then trained using the training dataset, while its performance was evaluated using the test dataset. Table 8 below shows a few evaluation metrics used to assess the performance of the decision tree classifier.

Table 8: Decision Tree Classifier Performance

Evaluation Metric	Value
Accuracy	87.8%
Precision	76.0%
Recall	77.6%
F1 Score	76.8%
AUC	93.0%

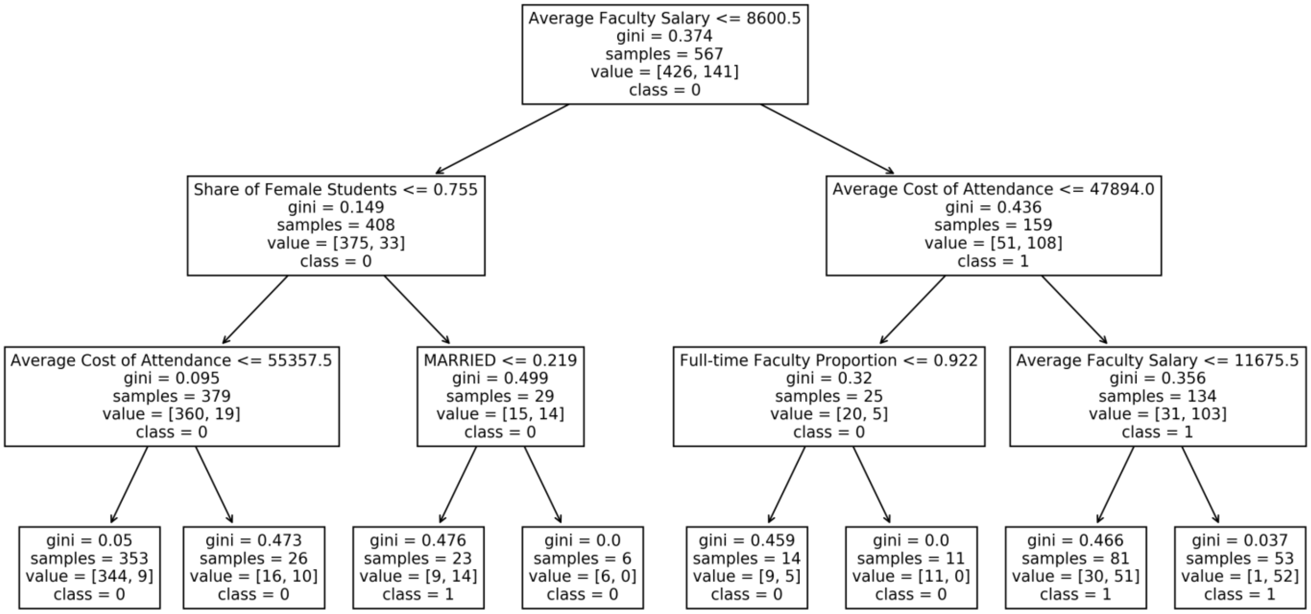
This model had the highest accuracy (87.8%) of all the models that were run, which means that it was the best in correctly labeling observations. It also has a high AUC value of 93%, which shows that the model is fairly good at distinguishing between the two classes. The most important features for this model are shown in Figure 8 below. These include average faculty salary, the average cost of attendance, and the share of female students.

Figure 8
Feature Importance



All three of these features are under an educational institution’s control i.e. educational institutions can change or modify all three in order to increase the probability of their students' earnings being in the top 25%. Looking at the decision tree in Figure 8 below, we can also derive more specific insights. We can see that when average faculty salary is less than or equal to $8601, the share of female students is greater than 75.5%, and the share of married students is less than or equal to 21.9%, student earnings are predicted to be in the top quartile.

Figure 9



VI. CONCLUSION

The insights discussed above have important implications for educational institutions since many institutional factors which are under an institution’s control are important predictors of students’ future earnings being in the top 25% across all colleges. This means that educational institutions can try to modify these factors in order to increase the probability of their students' earnings being in the top 25%. High student earnings not only improve the ranking of a college, but also increase admission applications since student earnings is one of the factors some students consider while deciding which colleges to apply to.

However, even though some important predictors of students' future earnings are under the control of educational institutions, these cannot be modified without proper procedures. For instance, with more and more colleges aiming for gender parity in their student body, systematically increasing or decreasing the share of female students can raise ethical and possibly legal concerns. Similarly, it would be unethical of educational institutions to filter applicants based on their household income in order to increase their median household income.

A limitation of the College Scorecard data is that the small number of public educational institutions as compared to private institutions do not allow a separate analysis to be conducted for only public institutions. Moreover, the dataset only records earnings for undergraduate Title IV recipients which raises concerns about how representative the results are of the entire student body. These limitations should be kept in mind while analyzing the results.

Future considerations include supplementing the College Scorecard data with other student and institutional characteristics. Moreover, a continuous instead of a binary target variable can be used in order to predict students' future earnings. Additional parametric and non-parametric techniques can also be employed e.g. using Random Forest to improve predictions.

BIBLIOGRAPHY

Cornwell, C. M., Mustard, D., & Parys, J. V. (2009). How Does the New SAT Predict Academic Achievement in College? *University of Georgia Working Paper*. https://www.researchgate.net/publication/228871851_How_Does_the_New_SAT_Predict_Academic_Achievement_in_College

Franklin University. (n.d.) Non-Profit vs. For-Profit Colleges: What You Need to Know. *Back To College Blog*. Retrieved April 16, 2021, from https://www.franklin.edu/blog/non-profit-vs-for-profit-colleges-what-you-need-to-know

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York Inc.

James, E., Alsalam, N., Conaty, J., & To, D. (1989). College Quality and Future Earnings: Where Should You Send Your Child to College? *The American Economic Review, 79(2),* 247-252. http://www.jstor.org/stable/1827765

King, G. & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis, 9,* 137–163. https://j.mp/2oSEnmf

Mashat, A. F., Fouad, M. M., Yu, P. S., & Gharib, T. F. (2012). A Decision Tree Classification Model for University Admission System. *International Journal of Advanced Computer*

Science and Applications, 3(10). https://www.researchgate.net/publication/235333385_A_Decision_Tree_Classification_Model_for_University_Admission_System

Miller, M. (2019). The Basics: Logistic Regression and Regularization. *Towards Data Science*. Retrieved May 14, 2021, from https://towardsdatascience.com/the-basics-logistic-regression-and-regularization-828b0d2d206c

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Pereira, J. M., Basto, M., & da Silva, A. F. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. *Procedia Economics and Finance, 39*, 634-641. https://doi.org/10.1016/S2212-5671(16)30310-0

Rumberger, R. W., Thomas S. L. (1993). The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review, 12(1)*, 1-19. https://doi.org/10.1016/0272-7757(93)90040-N

Strand, M., Truong, T. (2015). Predicting Student Earnings After College. *CS 229 Machine Learning Final Projects, Stanford University*. http://cs229.stanford.edu/projects2015.html

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58(1)*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

U. S. Department of Education. (2021a). College Scorecard Data. Office of Planning, Evaluation and Policy Development (OPEPD). https://collegescorecard.ed.gov/data/documentation/

U. S. Department of Education. (2021b). Technical Documentation: College Scorecard Institution-Level Data. Office of Planning, Evaluation and Policy Development (OPEPD). https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf

Watts, T. W. (2020). Academic Achievement and Economic Attainment: Reexamining Associations Between Test Scores and Long-Run Earnings. *SAGE Journals*, *6(2)*. https://doi.org/10.1177/2332858420928985

IMPLEMENTATION APPENDIX

The College Scorecard dataset was obtained through the College Scorecard API, and comprised of 6469 observations. Obtaining data through the API allowed me to obtain data only for currently operating educational institutions and to select only the variables of interest, which amounted to 21. These included student and institutional characteristics, the target variable, as well as variables needed to subset the dataset further e.g. a variable indicating whether or not an institution is a 4-year one was included in order to keep only 4-year institutions during the preprocessing stage. Many duplicate values (27% of the entire dataset) were found. These were present because not all multi-branch institutions report information needed to calculate the metrics “at the more granular branch location level” (U.S. Department of Education, 2021b), resulting in the same attribute values across different branches of the same institution. Therefore, only the first instance of the duplicate values was kept, while the rest were dropped from the dataset.

The dataset was reduced to include only those rows corresponding to currently operating, 4-year institutions. There were a lot of missing values in continuous predictors however, with the most missing values in a variable being 50%. Rows containing missing values of this variable were dropped, which significantly reduced missingness in the dataset and reduced the number of observations to 1209. The rest of the missing values were imputed using the means of the non-missing observations of each variable.