STAT 486 Final Project: Expected College Return on Investment

Tyler Ward

December 19, 2023

1 Introduction

Choosing a college to attend is no small feat, with countless different factors influencing the decision of a prospective student, and at least 1500 active 4-year degree awarding universities in the United States. Potentially foremost among these factors is the ”return on investment” (ROI) of students once they attend these universities; namely, how well education at a given university will prepare them for meaningful, competitively paying employment when their time at school is done. I explored the use of several machine learning models trained to estimate expected income of a students six and 10 years post entry to a given university. These models were trained on data made publicly available by the United States Department of Education (i.e., the College Scorecard). I used both simple and flexible models and considered their different strengths and weaknesses in order to find an optimal method that both could estimate expected income well in future years, fit the data well enough to feel confident about relationships learned by the model, and that would be interpretable enough to discuss how different variables contributed to expected six and 10 year post-entry income.

2 Exploratory Data Analysis

Exploring data from approximately 2000 four-year degree awarding universities helped me both see potential relationships that could effective help estimate mean income, such as several highly correlated variables like faculty salary, attendance cost, and SAT (or equivalent) score, as shown in the correlation matrix in Figure 1. Beyond these general relationships alone, exploratory analysis also helped show that in terms of earning, there appears to be a grouping of schools that fall high above the trends of expected earnings based on several variables even after splitting the observation by School ownership. Plots of earnings versus SAT average scores and attendance cost, as shown in Figure 2, show this group of schools clearly. After further examination of these observations, it became clear that these observations belonged to Ivy League schools, or Ivy Plus schools such as MIT, Stanford, and Duke. This prompted feature engineering to include this information in estimation by machine learning models.

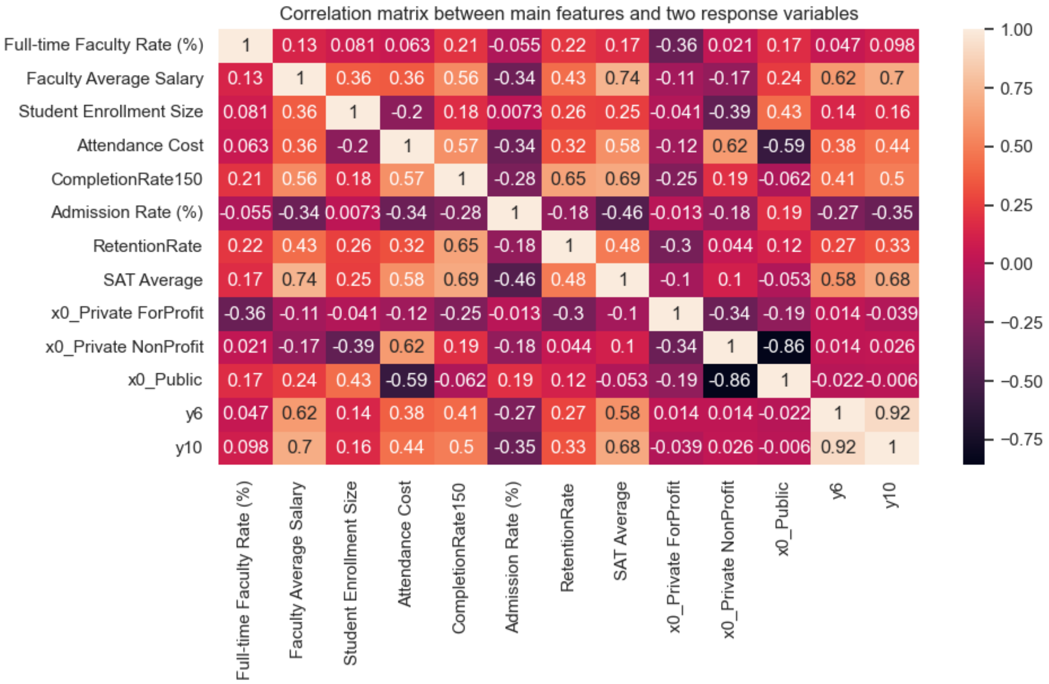
Additionally, several columns appeared to have sizeable amounts of missing values. This prompted the use of KNN Imputation to fill in missing numeric data as a preprocessing step before model training.

3 Machine learning models

A concise list of models attempted is found in Table 1. Metrics shown are approximate out of sample performance for 6-year post-entry income using on a 30 % test split after tuning with 5-fold CV. Scaled RMSE is the number of standard deviations away predictions were, on average, from observed values (therefore, lower values are better). XGBOOST is a method we did not discuss in this class, but which has many hyperparameters and boasts high performance and easy parallel computing, making it an important method in industry.

Key Hyperparameters explored for each of the models in Table 1 are listed below.

- LASSO: *alpha* parameter chosen via CV as weight to apply to each additional β in cost function



	Full-time Faculty Rate (%)	Faculty Average Salary	Student Enrollment Size	Attendance Cost	CompletionRate150	Admission Rate (%)	RetentionRate	SAT Average	x0_Private ForProfit	x0_Private NonProfit	x0_Public	y6	y10
Full-time Faculty Rate (%)	1	0.13	0.081	0.063	0.21	-0.055	0.22	0.17	-0.36	0.021	0.17	0.047	0.098
Faculty Average Salary	0.13	1	0.36	0.36	0.56	-0.34	0.43	0.74	-0.11	-0.17	0.24	0.62	0.7
Student Enrollment Size	0.081	0.36	1	-0.2	0.18	0.0073	0.26	0.25	-0.041	-0.39	0.43	0.14	0.16
Attendance Cost	0.063	0.36	-0.2	1	0.57	-0.34	0.32	0.58	-0.12	0.62	-0.59	0.38	0.44
CompletionRate150	0.21	0.56	0.18	0.57	1	-0.28	0.65	0.69	-0.25	0.19	-0.062	0.41	0.5
Admission Rate (%)	-0.055	-0.34	0.0073	-0.34	-0.28	1	-0.18	-0.46	-0.013	-0.18	0.19	-0.27	-0.35
RetentionRate	0.22	0.43	0.26	0.32	0.65	-0.18	1	0.48	-0.3	0.044	0.12	0.27	0.33
SAT Average	0.17	0.74	0.25	0.58	0.69	-0.46	0.48	1	-0.1	0.1	-0.053	0.58	0.68
x0_Private ForProfit	-0.36	-0.11	-0.041	-0.12	-0.25	-0.013	-0.3	-0.1	1	-0.34	-0.19	0.014	-0.039
x0_Private NonProfit	0.021	-0.17	-0.39	0.62	0.19	-0.18	0.044	0.1	-0.34	1	-0.86	0.014	0.026
x0_Public	0.17	0.24	0.43	-0.59	-0.062	0.19	0.12	-0.053	-0.19	-0.86	1	-0.022	-0.006
y6	0.047	0.62	0.14	0.38	0.41	-0.27	0.27	0.58	0.014	0.014	-0.022	1	0.92
y10	0.098	0.7	0.16	0.44	0.5	-0.35	0.33	0.68	-0.039	0.026	-0.006	0.92	1

Figure 1: Correlation matrix of several features with mean income 6 (y6) & 10 (y10) years post-entry

Table 1: Models attempted in analyzing expected post-entry income

Model	Description	Hyperparameters	Scaled RMSE	R^2
OLS	Ordinary Least Squares (Linear Regression)	N/A	0.74	0.45
LASSO	Penalized Regression removing variables leading to overfitting	one: α	0.76	0.43
KNN	k-nearest neighbors regression	1	0.74	0.45
RF	Bootstrapped and aggregated decision trees	4	**0.66**	**0.57**
XGBOOST	Extreme gradient boosting	3	0.67	0.56

- KNN: n_neighbors
- Random Forest (RF): n_estimators, max_features, min_samples_split, and min_samples_leaf
- XGBOOST: colsample_bytree, learning_rate, and max_depth

3.1 Model Selection

When attempting each of the models shown in Table 1 it was clear that ensemble models performed markedly better than single models, with the Random Forest and XGBOOST models consistently achieving lower RMSE results and higher R^2 results than the OLS, LASSO, and KNN non-ensemble models. KNN specifically struggled with extreme overfitting at many values of `n_neighbors`, achieving perfect in-sample fit, but no improvement in out-of-sample fit compared to linear regression, even after hyperparameter tuning. It is also interesting to see how a LASSO fit, though it restricts the model to only contain 3 features, still has a close to comparable fit to the linear regression model. It is clear that the OLS model struggled from multicollinearity present in the training data features, which could be seen by counter intuitive regression coefficients.

CATBoost was also considered, as I have heard it is a very effective method, but then I realized it is meant to deal with ”CAT”egorical features well, which would lend itself better for a classification focused analysis. I also considered using a Random Forest Quantile Regressor to be able to get intervals on expected income for each university, which may end up being something I implement into my dashboard of my results;

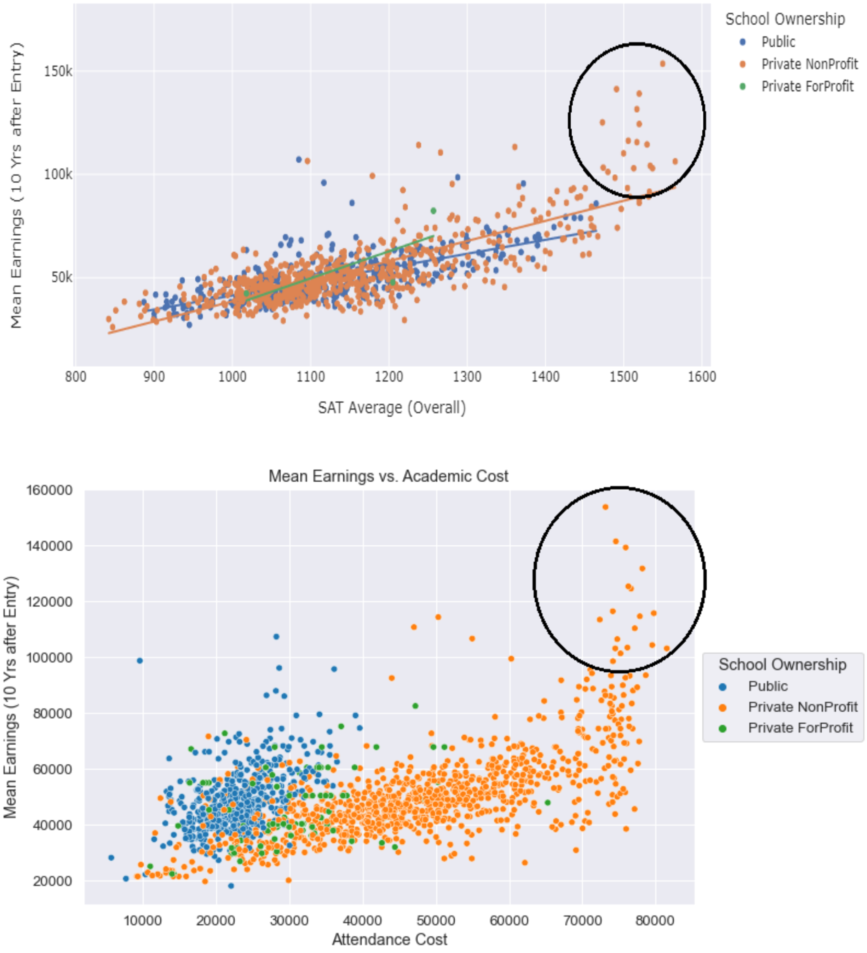


Figure 2: Plots showing clear grouping of High Earnings Universities (circled): Prestigious Ivy/Ivy Plus Universities

however, this model ended up being out of scope for the purpose of this project for now, with many moving pieces already in place.

3.2 Best Model

As shown in the high level results in Table 1, the model that performed best was the Random Forest model. After performing hyperparameter tuning for n_estimators, max_features, min_samples_split, and min_samples_leaf, it appeared that the best results in terms of RMSE came with max_features (the number of randomly selected features for each tree) set to the square root of the total number of feature, with the number of decision trees set to 250, and with min_samples_split, and min_samples_leaf set to their defaults of 2 and 1, respectively.

The `shap` package helped give greater understanding of why certain predictions are being made, and which variables contribute most to the predictions, by calculating the average contribution of changes in a given input feature to predictions. As shown in Figure 3, it is clear that faculty salary has the largest impact on expected income, and predictions appear to increase in a fairly linear fashion as faculty average salary increases. This is followed by SAT Average scores, completion rates, and attendance cost, when accounting for all other variables. Interestingly, before including the ”Ivy league plus” flag, attendance cost was the clear third most important variable. However, after including this new flag, it is approximately equivalent in average impact completion rate. Scatter plots of the SHAP values (also know as dependence plots) by the values of completion rates and attendance cost for each university are shown in Figure 5 and 6, respectively.

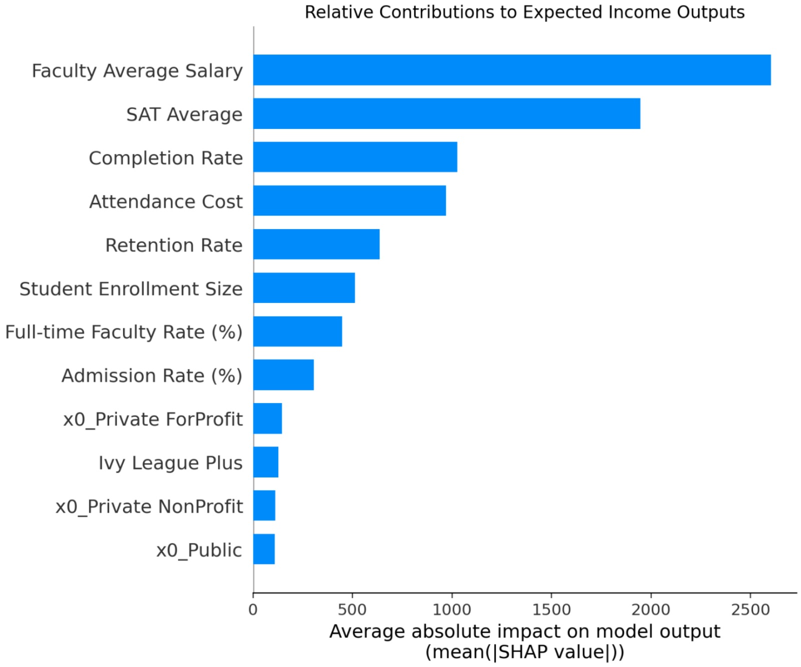


Figure 3: Importance of each variable to predicting expected income

This shows a benefit of including this flag in the model. It also shows the power of the implicit regularization applied by the Random Forest model, which is done as only a random subset of features are considered in each tree, and so if a variable is not as helpful towards minimizing the squared error of predictions, it will not be as influential on a given tree, and therefore will not be as influential in the final prediction (an average of the predictions of all 250 decision trees). This random sampling of features, along with bootstrapping and aggregating inherent helped random forest have the lowest out of sample RMSE and the highest R^2 value, and also helps random forest be robust against overfitting, a major benefit of this model over XGBOOST.

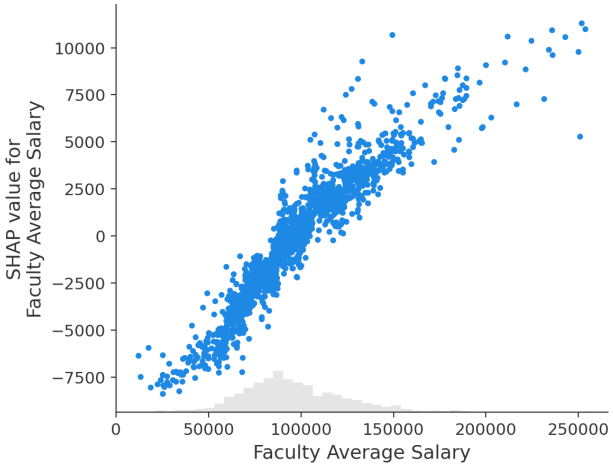


Figure 4: Contribution of Faculty Average Salary to Expected Income

4 Conclusion and Next Steps

A Random Forest model allowed for estimation of expected income of students 6-years post entry into a given university with the least error (RMSE) among both linear, single non-parametric, and ensemble models. It additionally was able to be fit with simplistic tuning, and a fair out of sample fit. Additionally, this model helped us understand relationship between different characteristics of colleges and expected income, with faculty average salary, average SAT scores, completion rates, and attendance cost have the largest absolute impact on predictions, on average.

Future work can be done to complete in-depth cluster analysis of the 4-year degree awarding colleges in my dataset, and then examine if this grouping of similar colleges can provide more information into expected income of alumni, and therefore improve model fit (R-squared) and predictive performance (RMSE). Additionally, with the amount of multicollinearity that exists in the dataset, exploring potential interactions in `shap` plots can help build greater interpretability of the model and the relationships between different characteristics of colleges and expected income.

A Supplementary Figures

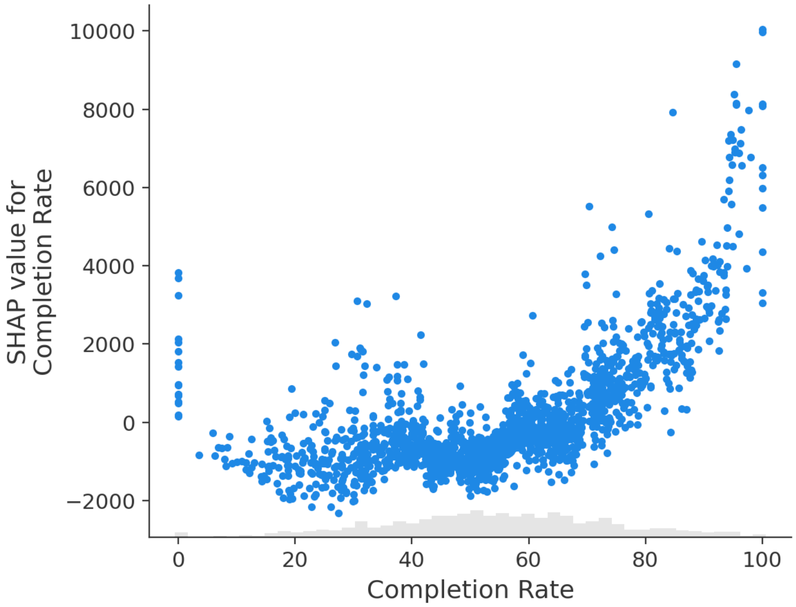


Figure 5: Dependence Scatter Plot of SHAP Values by School Completion Rate

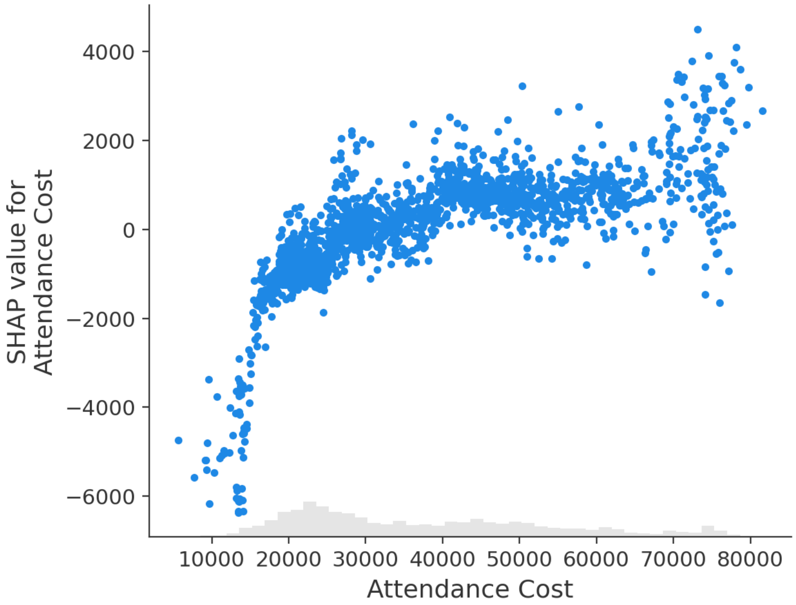


Figure 6: Dependence Scatter Plot of SHAP Values by Attendance Cost