Paycheck Pathways: Unveiling the Key Factors Shaping Early Career Earnings

Carter Hall * Connor McNeill * Miles Woollacott *

April 21, 2024

Abstract

This report analyzes factors that influence the early career pay of graduating college students beginning their careers. Through statistical techniques from linear regression to Bayesian Additive Regression Trees (BART), candidate models are fit in succession to create a narrative that highlights the modeling process. It was found that factors such as the average faculty salary, the student body's average SAT score, the percentage of the student body in STEM fields, and demographic proportions such as the percentage of students identifying as either Women or Asian were important factors in predicting early career pay. Through comparison of the number of predictors and the root-mean squared [prediction] error amongst the training and test sets, partial least squares regression and BART emerged as candidate models. Future work could involve the inclusion of student-centric data, focusing on universities' predominant majors and how the proportion of students in those majors might inflate/deflate early career pay figures.

Keywords: Salary, Higher Education, Stepwise Regression
Lasso, Partial Least Squares, Gamma GLM, BART

1 Introduction

Our motivation for this project is ultimately about what factors impact salaries of students post-graduation. Recent research conducted by Payscale showed that where students attended college was a key factor into how much money they made after 10 years in their career [Picchi, 2023]. However, this research did not get into the specific factors that caused this discrepancy in earnings. Knowing these factors may be critical to students when choosing which school to attend for their post-secondary education. But these some of these factors may not deal with aspects of the university and may be due to overall socioeconomic factors of the students at the university.

*Department of Statistics, North Carolina State University

Most of the top-ranked schools in the country are private schools with extremely cost of attendance. As such, even if students are accepted to these universities, they may not be able to attend due to the costs involved. Location is a factor to consider in this as well, due to two possible reasons: (1) out-of-state fees at public universities and (2) commuting to a local school allows students to save money. Additionally, some schools may have a higher percentage of students that are in majors that lead to in-demand high-paying jobs such as those in STEM fields. This leads to the question of what factors specifically lead to a higher-paying job after graduation.

Throughout this paper, we will utilize different modeling techniques to explore what factors about universities in the United States lead to differences in early career salary. The goal is that by using multiple modeling techniques, we may be able to identify factors that show up among multiple models. This would indicate that they are quite important for predicting students' salary as they graduate from college. Additionally, we will evaluate each of the models to determine which one does the best job of predicting early-career salary, and what predictors are utilized (or ignored) by each model.

2 The Data

2.1 Data Sources

The data utilized in this project originally came as a few datasets which we combined. This combination comes naturally since ultimately all the data comes from publicized data from the United States Department of Education (DOE). The primary datasets come from Kaggle [Mostipak, 2020] which included three smaller datasets. Additionally, we obtained data from the DOE's College Scorecard [U.S. Department of Education, 2023] in order to get more up-to-date detailed information about each school which the Kaggle data did not provide.

The Department of Education publishes the College Scorecard data in order to provide "increased transparency [into] how well individual post-secondary institutions are preparing their students to be successful; [helping] students and their families compare college costs and outcomes as they weight trade-offs of different colleges" and the needs of the student [U.S. Department of Education, 2023]. The dataset itself is quite messy to work with, which is why it was used solely as a means to add additional covariates to consider in our modeling. But having data directly from the DOE is extremely useful in our analysis.

In summary, the following are the sources used for data:

- Tuition and fees along with diversity information from O'Leary and Hatch [2016] and The Chronicle of Higher Education [2021] (via Kaggle)
- Salary potential data from Payscale [2023] (via Kaggle)
- College-specific data on admissions, socioeconomic status, and financial aid from College Scorecard [U.S. Department of Education, 2023]

Since we are evaluating our models based on prediction, splitting our data into a training set and a test set is key in order to avoid overfitting. Our sample size is $n = 683$ colleges, so we will utilize a 80% training/20% test set split in order to have enough data to have good model fits that are trained on a majority of the data and also be able to check for model performance without an overfitting bias.

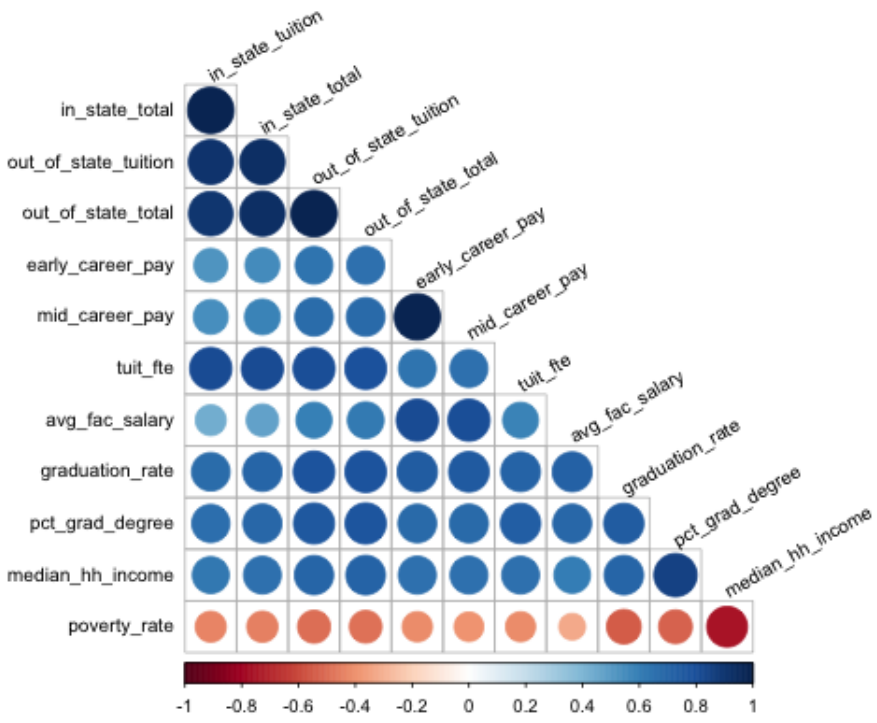

Figure 1: This table shows a subset of the correlations between the variables in our dataset. Variables that had at least one correlation of $|r| > 0.75$ were included.

2.2 Linear Dependencies & Addressing Multicollinearity

Before fitting any models and exploring our dataset, a few variables were removed from consideration as they did not provide any value as predictors – the name of the college or university, the state in which the school is located (due to having too many factors, as well as being represented by predictors for the region of a school), and the mid-career pay (as utilizing a predictor obtained

after the recording of the response does not make sense).

The immediate problem that arose with our modeling techniques were the linear dependencies that existed among some of our predictors. These linear dependencies came up as some of our predictor variables were linear combinations of other predictors. We removed each linear dependency from our models by dropping the dependent variables until we had no linear dependencies.

A matrix of select correlations between our predictors can be found in Figure 1. The correlation matrix indicates a high number of correlated predictors. This is not surprising, based on the context around our data. For instance, the total cost of attendance for in-state students includes in-state tuition, which are two separate variables. While not exact linear combinations, this likely will become an issue with models where multicollinearity is an issue.

In order to explore whether there is an issue of multicollinearity within our data, we need to fit a linear regression model. Employing this model is a natural step in our exploratory data analysis. The first goal is to identify and handle predictors with a Variance Inflation Factor (VIF) larger than 10, a threshold that signals potential issues related to multicollinearity, and remove them from consideration in further modeling.

From this, repeated consultation of linear regression results and subsequent removal of predictors with extreme VIF values or linear dependencies eliminated the following predictors from consideration (the table of extreme VIFs after linear dependencies removed can be found in Table 6):

- The total percentage of students who come from designated minority groups along with the percentage of Non-resident Foreign students (as these two were already represented in the model) and the percentage of African American students (due to strong correlation with whether the school was an HBCU or PBI).
- The type of degree that the majority of students graduate with and highest degree the university offers (which were correlated with the length of degree variable)
- The out-of-state and in-state cost to attend a school, including both tuition and room & board (notably correlated with our out-of-state and in-state tuition covariates, respectively).
- The state-wide rank of the school's alumni salary earnings (as we have variables in our dataset such as region and early career pay that already measure this)
- The percentage of students at the university who's family's income puts them below the poverty line (correlated with a number of socioeconomic covariates).

2.3 Variable Transformations & Influential Points

Figure 2 is a histogram of our response variable, early career pay, for every college we had data on. Note how the histogram indicates that our response variable only has positive values, and is

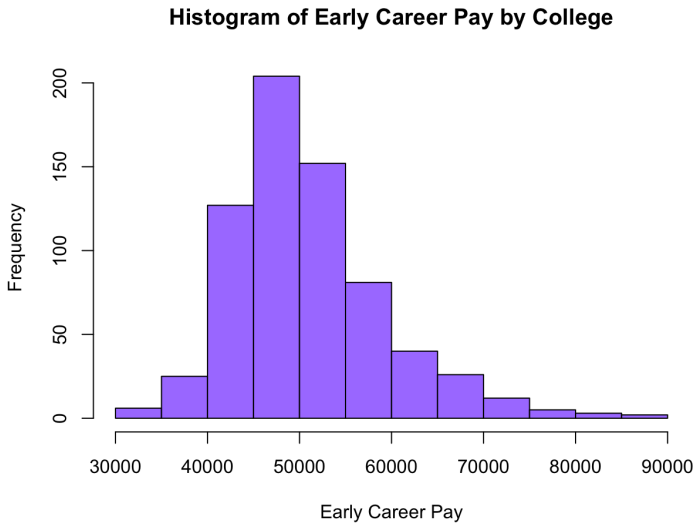

Figure 2: A histogram of early career pay for all colleges in our dataset.

right-skewed. This indicates that we may need to transform our response variable, or fit a model that accounts for a purely positive response variable.

The previous procedure introduced a baseline linear regression model through which the necessary predictor and response transformations were evaluated. A Box-Cox transformation [Box and Cox, 1964] was consulted, informing a log-transformation of our response after obtaining $\lambda = 0.101$. This transformation has, in fact, been studied by Ganzach and Pazy [2021] under similar contexts regarding salary data, and is reflected by the empirical distribution of the response variable shown earlier.

After log-transforming the response variable, a Box-Tidwell transformation was considered for a number of covariates which had solely nonzero values. This resulted in transformations for total enrollment (logarithmic), admission rate (inverse), and the proportion of domestic students (quartic).

One outlier was removed from the training dataset after transformations was due having the largest standardized residual value (6.0054) in our dataset. While there were two more influential points that could potentially be classified as outliers, we decided not to remove them as they did not impact our overall model much. Additionally, more linear dependencies appeared once these observations were removed and model fitting was attempted.

3 Modeling

3.1 Linear Regression Methods

3.1.1 Stepwise Regression

With a potentially useful predictors in our dataset, we thought it prudent to determine which subset of predictors accurately maximized predictive capabilities, while minimizing the risk of overfitting. Therefore, we turned to stepwise linear regression, using the variable transformations we established in our initial linear regression model, using 10-fold Cross-Validation to select the number of predictors.

Coefficients for our final model are displayed in Table 1, along with diagnostic plots in Figure 4. Stepwise regression successfully reduced the number of predictors within our model from 42 to 7, while still fitting our data rather well, with $R^2_{\text{adj.}} \approx 0.8224$, and a RMSEP ≈ 0.0775.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.5680	0.1780	59.37	< 0.0001
Avg SAT Score	-1.564e-04	3.861e-05	-4.05	0.0001
% Students in STEM	0.3241	0.0205	15.79	< 0.0001
Tuition Revenue per Student	3.310e-06	4.807e-07	6.89	< 0.0001
Avg Faculty Salary	2.256e-05	1.707e-06	13.21	< 0.0001
% Students with Pell Grants	-0.2935	0.0268	-10.95	< 0.0001
% Domestic students	0.5419	0.2727	1.99	0.0474
(% Domestic)^4	-0.3730	0.1068	-3.49	0.0005

Table 1: Coefficients from our final model from stepwise regression, along with individual test statistics and p-values.

Increasing the SAT average scores is related with a decrease in the early career pay. There is no logical explanation for this; mathematically, however, the coefficient on the log-scale is nearly zero – transforming this to the original scale, we obtain a value of ≈ 0.999, meaning a very *slight* decrease in career pay for an increase in mean SAT score.

Based on our model coefficients, increasing the percentage of STEM students is related with an increase in early career pay. This isn't surprising, as STEM graduates generally pursue careers in medicine and technology, amongst other fields.

It initially seems like an increase the university expenditure on students is correlated with an increase in early career pay. However, we know from EDA that there exists a correlation with universities' expenditure on students and tuition. Therefore, this trend could be due to tuition costs themselves; investigating tuition while holding other predictors fixed, we see a 3% increase in expected salary per $10000 increase in tuition.

Interestingly, an increase in the percentage of students on Pell Grants is related with a $\approx$ 26% decrease in expected early career pay. This is possibly due to some universities being located in lower-income areas where jobs do not fetch as high a salary as compared to more major cities, or from students using a degree as a means to transition to another area altogether.

An increase in the percentage of domestic students is related with an increase in early career pay. Given the presence of higher-order polynomial terms, we could analyze the increase in response per unit increase by $\exp\{0.5419(d+0.01) - 0.3730(d+0.01)^4 - (0.5419d - 0.3730d^4)\}$, if we denote the percent of domestic students by $d \in [0, 1]$. This relationship is concave in d, albeit slightly, meaning that an increase in the proportion of domestic students decreases expected early-career pay slightly.

While our model predicts relatively well and has a small number of relevant predictors, the slightly negative coefficient of average SAT scores is peculiar. In addition, one could argue the presence of fanning residuals in Figure 4, although this perhaps is too critical. Thus, in searching for an interpretable model that makes intuitive sense, we introduce another variable-selection technique – the LASSO.

3.1.2 LASSO

A $k = 10$-fold cross validation approach was used to fit a LASSO model. Figure 5 is a plot of mean squared error (MSE) versus various choices of λ. The plot indicates that the choice of $\lambda = 0.00627$ fits our data well enough (an MSE within one standard error of our empirical minimum MSE for all λ values), and is as large as possible.

Table 2 lists the coefficients for Lasso Regression at our optimal value of λ. This model included more terms than stepwise selection; however, the final models under these approaches were not nested, meaning an analysis-of-deviance test would be improper to test sufficiency of the LASSO.

Variable	Coefficient	Variable	Coefficient
(Intercept)	10.7034	% Students with Pell Grants	-0.1957
Graduation Rate	0.0826	% Students who "Better World"	0.0391
% Students in STEM	0.2263	% Parents with Graduate Degrees	0.2248
School in Remote Town	-0.0110	(% Domestic)^4	-0.1143
Tuition Revenue per Student	1.769e-06	% Students identifying as Female	-0.0930
Avg Faculty Salary	1.761e-05	% Students identifying as Asian	0.0979

Table 2: Lasso coefficients for $\lambda = 0.00627$

3.1.3 Partial Least Squares

We also wanted to explore a model that includes all of our candidate predictors, but still performs dimension reduction in a different way. As such, we consider PLS with 10-fold Cross-Validation to select the number of components.

A plot showing the RMSEP versus the number of components is displayed in Figure 6. The smallest number of components that fits our data "reasonably well" (within one standard error of the minimum RMSEP against all possible components) is 3, so we keep the top three components.

	Component 1	Component 2	Component 3
% Students in STEM	0.0095	0.0264	0.0369
log(total enrollment)	0.0046	0.0148	0.0127
Tuition Revenue per Student	0.0103	0.0088	0.0145
Avg Faculty Salary	0.0128	0.0201	0.0233
% Students on Pell Grants	-0.0097	-0.0174	-0.0232
Graduation Rate	0.0119	0.0137	0.0165
% Households with Graduate Degree	0.0111	0.0093	0.0104
% Students identifying as Female	-0.0061	-0.0202	-0.0217
% Students identifying as Asian	0.0095	0.0128	0.0147

Table 3: Table of loadings of select predictors for the three loadings in the final PLS model. Predictors were included if there was a loading α that satisfied $|\alpha| > 0.015$ for any of the three components.

Table 3 displays the predictors with the highest loadings in any of the three components. Most of these predictors are the same as those in Stepwise Regression. However, we note that the percent of domestic students and the average SAT score variables don't contribute much to the components. On the other hand, total enrollment, graduation rate, percentage of graduate students, and the percentage of students identify as female or Asian contribute more to the components than what was indicated in stepwise selection.

3.2 Gamma GLM

In the previous section, we explored models that met the necessary assumptions to fit an ordinary least squares regression. However, in order to not violate these assumptions, a log-transformation was needed on our response variable, early-career pay. As such, it is worth the time to consider a generalized linear model (GLM) that would allow us to fit a model without needing to transform the response. Since salary pay is financial data, we know the response to be non-negative and relatively high in magnitude, which is a sign that our errors are not actually normally distributed. In fact, using a log-link in our Gamma GLM is somewhat of a similar approach to Gaussian model with a log-response when the variance is small. As such, we will utilize a log-link in our Gamma

GLM model. To summarize this mathematically, the link function is $\eta = \log(\mu)$. Figure 8 shows that the log-link is an appropriate link function to choose.

Additionally, we reparameterize the Gamma distribution slightly when fitting a GLM. Instead of the typical shape-scale parameterization, the shape parameter ν remains the same while we represent the scale parameter as $\lambda = \nu/\mu$ since $\mu = \nu\lambda$. The parameterization is useful since our link function is a function of μ. This gives us the following density function:

$$f(y;\nu,\mu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left\{-\frac{y\nu}{\mu}\right\}, y > 0$$

We will need to utilize some sort of variable selection screening to fit the Gamma GLM in order to overfitting. As such, we will utilize the variables selected by our Stepwise linear model from Section 3.1.1. This means that if our variance is small (i.e. the dispersion parameter is small since it is proposed by McCullagh and Nelder [1989, pg. 285-293] that $\hat{\phi} = \hat{\nu}^{-1}$), then our results of the two models will be extremely similar since the log-link function is used [Faraway, 2016, pg. 177].

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.5709	0.1794	58.91	<0.0001
Avg SAT Score	-1.568e-04	3.892e-05	-4.03	0.0001
% Students in STEM	0.3243	0.0207	15.67	<0.0001
Tuition Revenue per Student	3.307e-06	4.845e-07	6.83	<0.0001
Avg Faculty Salary	2.233e-05	1.721e-06	12.97	<0.0001
% Students with Pell Grants	-0.2939	0.0270	-10.87	<0.0001
% Domestic Students	0.5474	0.2749	1.99	0.0469
(% Domestic)^4	-0.3778	0.1077	-3.51	0.0005

Table 4: Coefficient estimates and individual tests of significance for final Gamma GLM

Table 4 shows the coefficient estimates along with their standard errors, and their individual tests of significance. First, we need to perform an Analysis of Deviance test to ensure that the model does a better job fitting the data than the null model. (Recall D is calculated as the difference between the deviance of a null model, i.e. a model with only an intercept, and the prospective model.)

$$D = 13.0257 - 2.2325 = 10.7932$$

It follows that $P(\chi^2_7 > D) = 0.1479$. As such, we fail to reject the null hypothesis. This means that the Gamma GLM does not do a good enough job of model fitting when compared to the null model. This is interesting though since the model is almost exactly the same as our linear regression

model after stepwise selection which had a very significant global F-test. Additionally, all of the p-values are significant in terms of our coefficients. Also, our plot of transformed fitted values vs. the deviance residuals look decent as well per Figure 7. The issue here with model fit also is seen with the AIC being 10371, much larger than what we observed for other models.

Our dispersion parameter is estimated to be $\hat{\phi} = 0.0042$. This means that we have a large value of ν, which indicates that our method should be very similar to utilizing a normal distribution with a log-transformed response [Faraway, 2016, pg. 175-177]. The coefficients, standard errors, test statistics, p-values, and the RMSE/RMSEP all are nearly the same. Even though the analysis of deviance test failed to reject the null hypothesis, there may be one benefit with working with the Gamma GLM over the linear regression model with stepwise selection: our response is not transformed, meaning that any model predictions do not need to be transformed.

3.3 Bayesian Additive Regression Tree (BART)

3.3.1 Summary of BART

In all of the models considered with the exception of PLSR, we assume the errors follow a known distribution, e.g., Gaussian or Gamma. However, a common theme underlying the previously explored approaches is the search for a *single* entity or model structure that, ideally, accurately explains the relationship between predictors and response.

BART, by contrast, is a tree-based ensemble method – a regression technique that combines multiple predictions to produce a more robust, and hopefully accurate, prediction. Some common ensemble methods include boosting, bagging, and random forests. The idea behind tree-based methods is to approximate some function of covariates, f, that predict a response, Y, through a *sum-of-trees model*

$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \tag{1}$$

where T_j is a binary regression tree, M_j is corresponding terminal node parameters, and ϵ is the error term assumed to be Gaussian with constant variance [Chipman et al., 2010, pg. 2]. With $x \in \mathbb{R}^p$ allowed to be multivariate, BART thus has the capacity to incorporate interaction effects between covariates. To do this, BART uses a sum-of-trees model similar to (1) and imposes a regularization prior on each of the (T_j, M_j) and variance σ^2 to shrink the effects of the individual trees. This makes each regression tree a so-called 'weak learner' of the response. Markov-Chain Monte Carlo (MCMC) methods are employed to iteratively update estimates.

By weakening the effects of individual trees, BART circumvents a problem of having one or few predictors dominate all of the predictions. While capable of variable selection and calculating variable importance, BART also has the ability to set a hyperparameter controlling the number of trees per iteration to a large value, which enables the model to investigate more nuanced relationships through variability introduced in additional trees.

Recent work by Kapelner and Bleich [2016] and Inglis et al. [2024] in developing R packages bartMachine and bartMan, respectively, have enabled the fitting of BART models with cross validation and succinct analysis and remarkable visualizations. Relevant hyperparameters include the number of trees, m, and mean/variance estimates for priors. (Our analysis focuses only on the *first* hyperparameter, as Kapelner and Bleich [2016, pg. 4] indicate that default values for others generally provide good performance.) It should be noted that the analysis involving BART was rather computationally expensive, with the former package requiring a Java backend. Memory challenges were encountered when using original hyperparameter values for the number of trees, hence the change to those listed below.

3.3.2 Fitting of a BART Model

To aid convergence of estimates in the MCMC process, 1000 burn-in iterations (iterations of estimates that occur but are then discarded) were specified in the fitting of a BART model through 4000 iterations after the burn-in period.

Table 5 presents the 10 best hyperparameter combinations with respect to (average) out-of-sample error (OOS Error) calculated on each fold. The final BART model fit was parameterized with $k = 5$ fold cross-validation on $m = 40$ trees per iteration (with other candidate values for m being $5, 10,$ and 20), with the aforementioned burn-in and remaining iterations. Figure 9 presents convergence-focused diagnostics; of-interest is the top-left plot showing the σ^2 estimates settling in-between the shown 95% credible interval.

As for checking relevant assumptions, a familiar assumption arises in Gaussian and homoskedastic errors. Figure 10 presents visualizations for checking these assumptions with an accompanying Shapiro-Wilk test, where we note the assumptions appear to be satisfied. Lastly, Figure 11 illustrates the overall strong performance of the BART model on the test dataset with accompanying credible intervals for each prediction.

3.3.3 Visualization of Tree Fits

Tree-based methods often come equipped with functionality to determine the relative variable importance – how useful a covariate was in the prediction of the response. BART is no exception, although it provides this relationship at both a predictor and tree-based level.

Figure 3 presents the variable importance plot for the BART model after training. We see similar predictors to many of the LASSO, PLSR, and GLM modeling approaches such as average faculty salary, percent of students in STEM, among others. This is corroborated in Figure 12 where a joint visualization between the most prominent trees and all trees amongst the final iteration of training are visualized. Notice that many of these trees are quite 'shallow' in terms of depth, with many reaching a leaf/node (called *Stump/Leaf* in the figure) after just 1-2 branches. The range of depth depicted in Figure 12 highlights the utility of BART to capture simple and complex relationships between covariates.

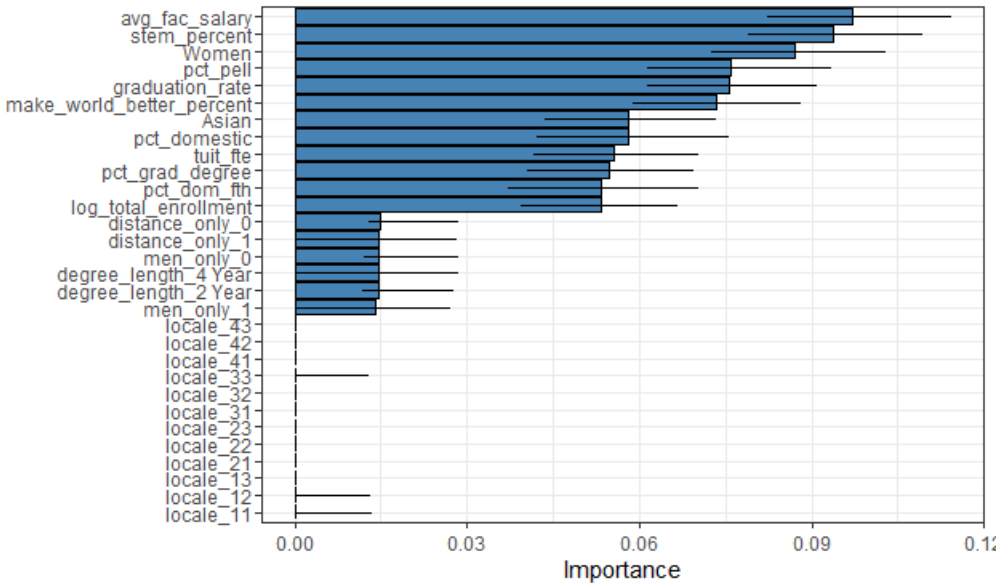

Figure 3: Variable importance plot for final BART model with predictor on the y-axis.

4 Results and Discussion

4.1 Model Comparison

Table 5 presents a summary of our models' performances under the root mean squared error (RMSE) criterion and number of non-zero coefficients (if applicable, including the intercept).

OLS performed better than stepwise regression on the training data, although this could be due to overfitting, as evidenced by the large difference in training and test errors. Similarly, stepwise regression is a simpler, more parsimonious model, with just one-sixth the coefficients of OLS, and a roughly equivalent accuracy on the test set.

LASSO posted metrics in-between OLS and stepwise, notably obtaining similar performance on the test set to its variable-selection counterpart with nearly double the coefficients. The Gamma model performed exactly the same as stepwise (due to reasons explained earlier), and we did not have to transform our response variable. However, we found that this model performed about as well as a log-transformed stepwise model.

In addition, the Gamma GLM may be difficult to interpret than a log-transformed model. PLS had the second-best testing RMSE and a third- training RMSE, an indication of a very good model fit. Furthermore, its dimension-reduction capabilities projecting a complex predictor set onto just 3 components underscores this point.

	OLS	Stepwise	Lasso	PLS	Gamma GLM	BART
RMSE (Train)	0.0557	0.0639	0.0634	0.0605	0.0639	0.0465
RMSE (Test)	0.0760	0.0775	0.0775	0.0726	0.0774	0.0655
Difference	0.0203	0.0136	0.0141	0.0121	0.0135	0.0190
Number of Nonzero Terms	53	8	15	3*	8	–

* number of components retained

Table 5: Summary of various comparison methods for our models. Note the errors are presented on the log-scale.

However, BART does the best job per the RMSE on both the training and testing sets. While its RMSE metrics on both training and testing were the minimum across the models considered, it notched the second-highest disparity (noted in the *Difference* column in Table 5) between the two. Overfitting could be the cause, although Chipman et al. [2010, pg. 4] detail recommended choices of hyperparameters to attempt to mitigate any overfitting which were followed in our modeling approach.

In choosing a final model, we highlight both PLS and BART for two reasons: the former due to its relative simplicity (having only 3 components) and speed compared to BART, and the latter for doing the best job at predicting while also giving us good, interpretable output regarding variable importance.

4.2 Conclusions and Future Work

We should note that, based on how we collected our data, we can only determine correlation, and not causation. Many of the associations found with early-career were common across the modeling techniques explored, such as average faculty salary and the proportion of students in STEM fields. Interestingly, the proportion of students identifying as female was often an important predictor in a model, although this cannot be thought of as fluctuations in this proportion causing a wage gap. Financial factors such as tuition revenue and average faculty salary along with socioeconomic factors such as the percentage of students receiving Pell Grants came up as important predictors in every model. This gives us evidence that students' financial and socioeconomic background may carry some weight in salary earnings upon graduation. However, we must clarify that our observations are characteristics of the overall student body at the school and not of the individual students themselves.

Beyond investigating what factors relate to increases in early-career pay, future analysis seeks to explore what happens to important factors in future earnings as time passes from graduation and students gain more work experience. Our dataset is naturally university-specific while predictions are for early career pay, a student or graduate-centric quantity. Thus, we might extend our analysis by obtaining records on the predominant majors of students at each university and the effect of such

demographics on salary. Interestingly, none of the predictors encoding locale, region, or state were found to be instrumental in explaining fluctuations in early career pay across the models considered. We could potentially extend to network analysis or a spatial modeling procedure to investigate the role of proximity between schools and/or their distance to notably high-paying major cities.

Bibliography

G. E. P. Box and D. R. Cox. An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964. ISSN 00359246. doi: 10.1111/j.2517-6161.1964.tb00553.x.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, Mar. 2010. ISSN 19326157, 19417330. doi: 10.1214/09-AOAS285.

J. J. Faraway. *Extending the Linear Model with R : Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC Press, Taylor & Francis Group, Boca Raton, 2016. ISBN 978-1-4987-2098-4.

Y. Ganzach and A. Pazy. The Scaling and Modeling of Pay and the Robustness of the Effect of Core Self Evaluations on Career Success. *Frontiers in psychology*, 12:608858, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.608858.

A. Inglis, A. Parnell, and C. Hurley. Visualisations for Bayesian Additive Regression Trees. *Journal of Data Science, Statistics, and Visualisation*, 4(1), Feb. 2024. doi: 10.52933/jdssv.v4i1.79.

A. Kapelner and J. Bleich. bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software*, 70(4):1–40, Apr. 2016. doi: 10.18637/jss.v070.i04.

P. P. McCullagh, 1952 and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989. ISBN 0-412-31760-5.

J. Mostipak. College tuition, diversity, and pay, Mar. 2020.

B. O'Leary and J. Hatch. Student Diversity at More Than 4,600 Institutions. https://www.chronicle.com/article/student-diversity-at-more-than-4-600-institutions/, Sept. 2016.

Payscale. Best Schools by Salary Potential. https://www.payscale.com/college-salary-report/best-schools-by-state, 2023.

A. Picchi. A college degree can boost your pay — but so can your alma mater. Here are top colleges for income. https://www.cbsnews.com/news/college-best-and-worst-colleges-impact-on-income-payscale/, Sept. 2023.

The Chronicle of Higher Education. Tuition and Fees, 1998-99 Through 2020-21. https://www.chronicle.com/article/tuition-and-fees-1998-99-through-2018-19/, May 2021.

U.S. Department of Education. College Scorecard, Oct. 2023.

5 Appendix

5.1 Supplemental Tables & Figures

Predictor	VIF
In-State Tuition	183.49
In-State Total	175.80
Region	88.28
Predominant Undergraduate Degree Type	27.74
% Students who identify as Black	26.71
Out-of-State Tuition	21.08
Predominant (General) Degree Type	20.88
Median Household Income	18.97
% Students who identify as White	15.97
Poverty Rate	14.90
% Households with Graduate Degree	11.38
Type of School	10.06

Table 6: This table shows the variance inflation factors (VIFs) that were greater than 10.

k	ν	q	m	OOS Error
5	3	0.90	40	0.0642
5	3	0.99	40	0.0647
3	10	0.75	40	0.0648
5	10	0.75	40	0.0648
3	3	0.90	40	0.0649
3	3	0.90	20	0.0653
2	3	0.99	40	0.0657
2	3	0.90	40	0.0657
5	3	0.90	20	0.0659
3	3	0.99	40	0.0663

Table 7: Top 10 Hyperparameter Combinations for Fitting BART Model on Training Data

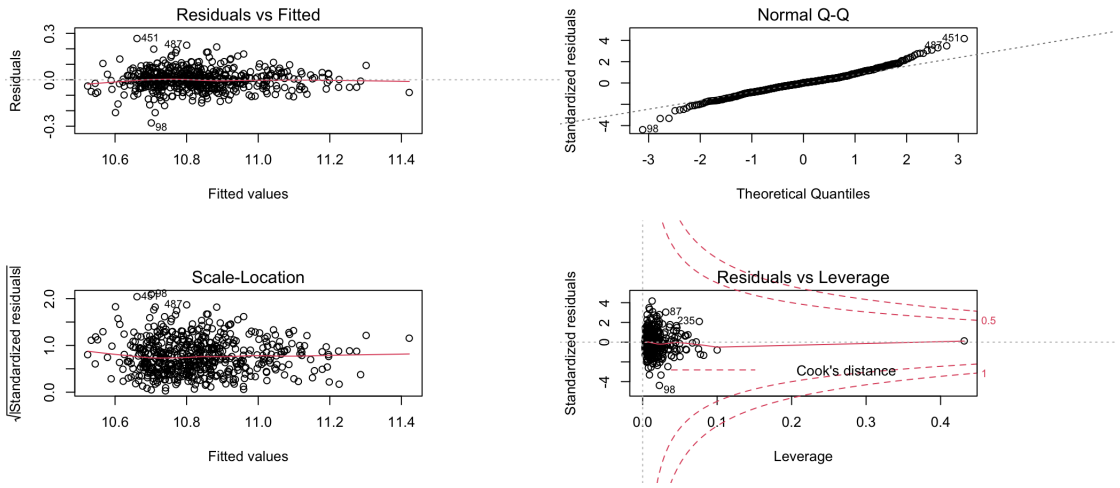

Figure 4: Residual plot from our final model using stepwise regression.

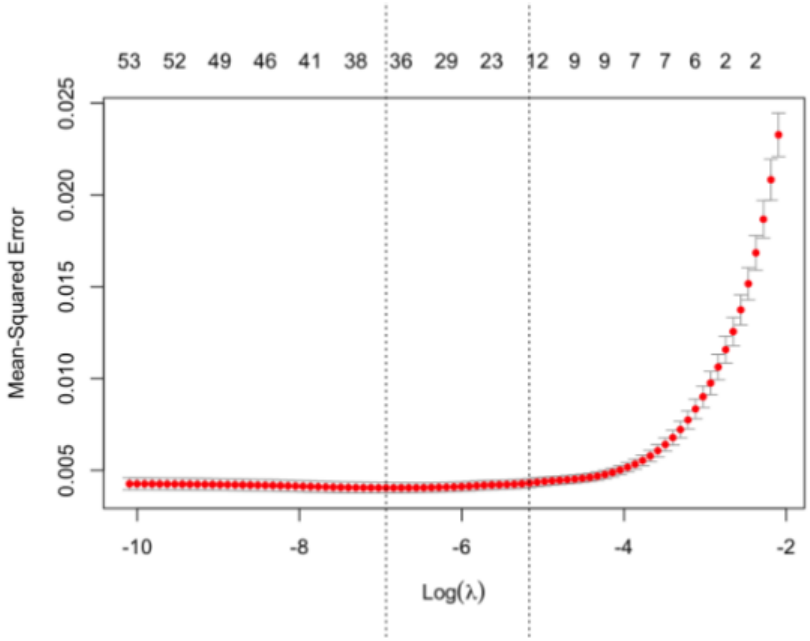

Figure 5: MSE versus the choice of λ.

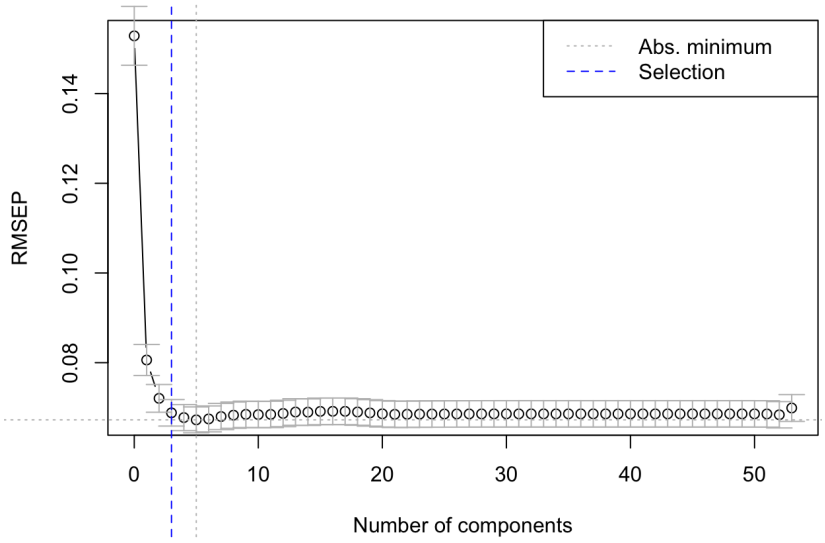

Figure 6: RMSE vs. Number of Components for PLS

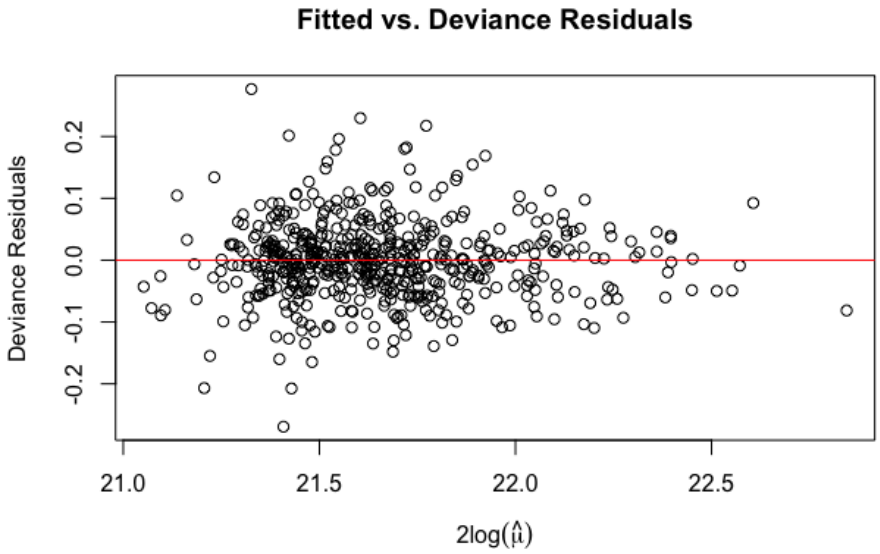

Figure 7: This figure shows our fitted values (on the log-scale multiplied by two as recommended by McCullagh and Nelder [1989]) vs. the deviance residuals. We see overall random scatter, which indicates a good model fit.

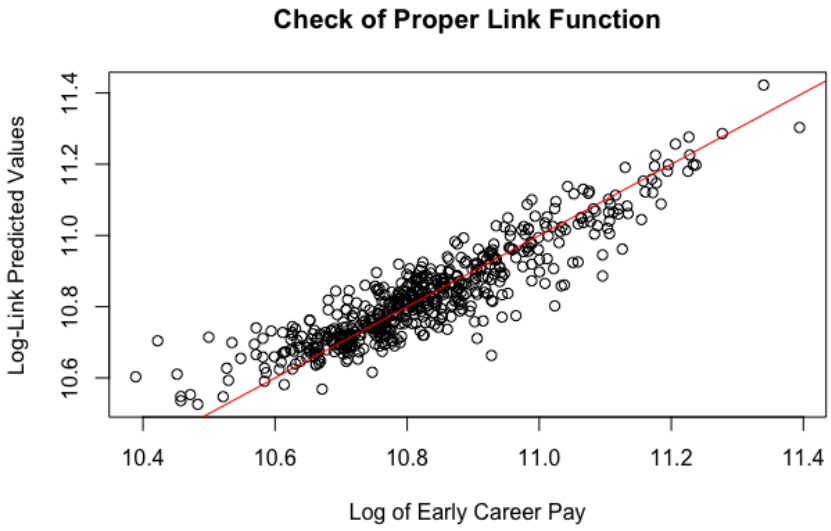

Figure 8: This figure shows the predicted log-link values vs. our transformed response paper, a check for proper residuals recommended by McCullagh and Nelder [1989].

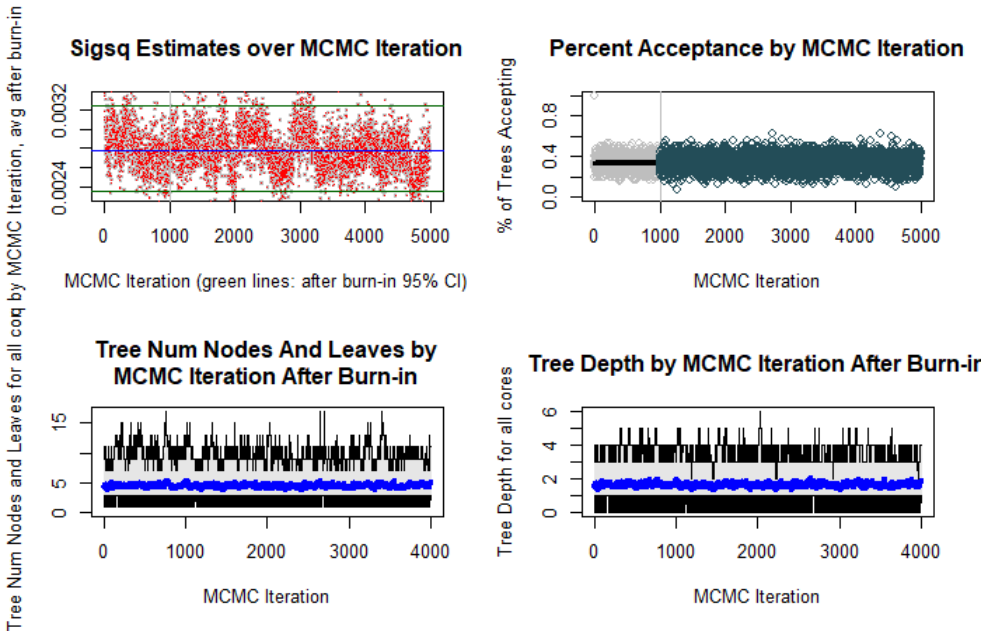

Figure 9: Convergence diagnostics for final BART model, including convergence of σ^2 estimates (top-left), percent acceptance via Metropolis-Hastings (top-right), numbers of leaves/nodes in trees (bottom-left), and depth of tree fits (bottom-right).

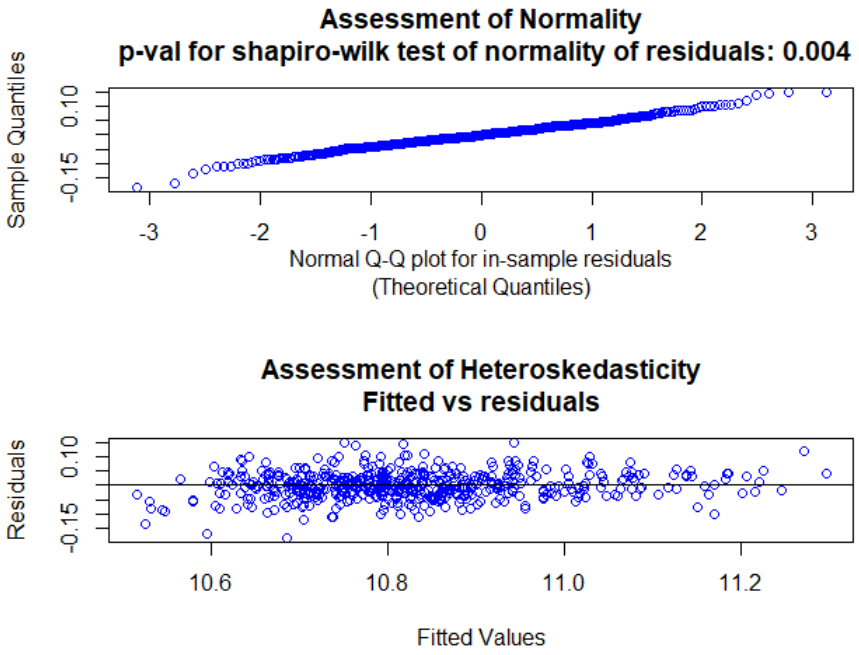

Figure 10: Normal Q-Q Plot and Fitted Values versus Residuals Plot for BART Model

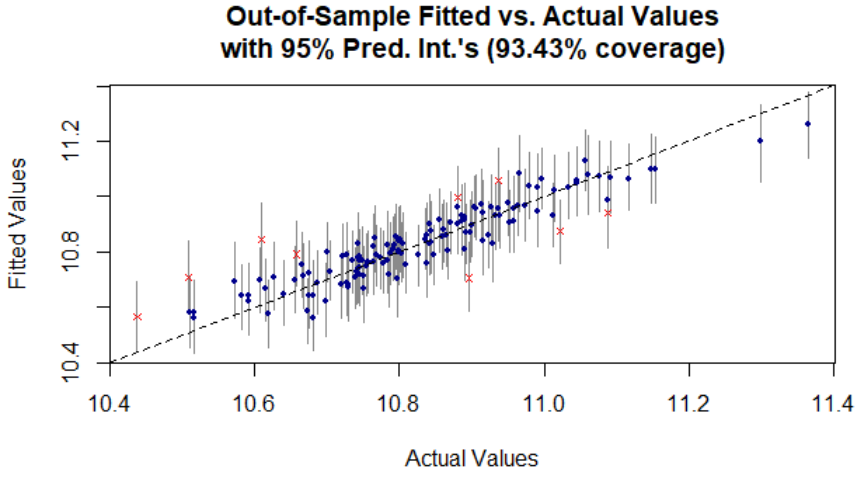

Figure 11: Fitted versus Predicted Values in Test Set for BART Model

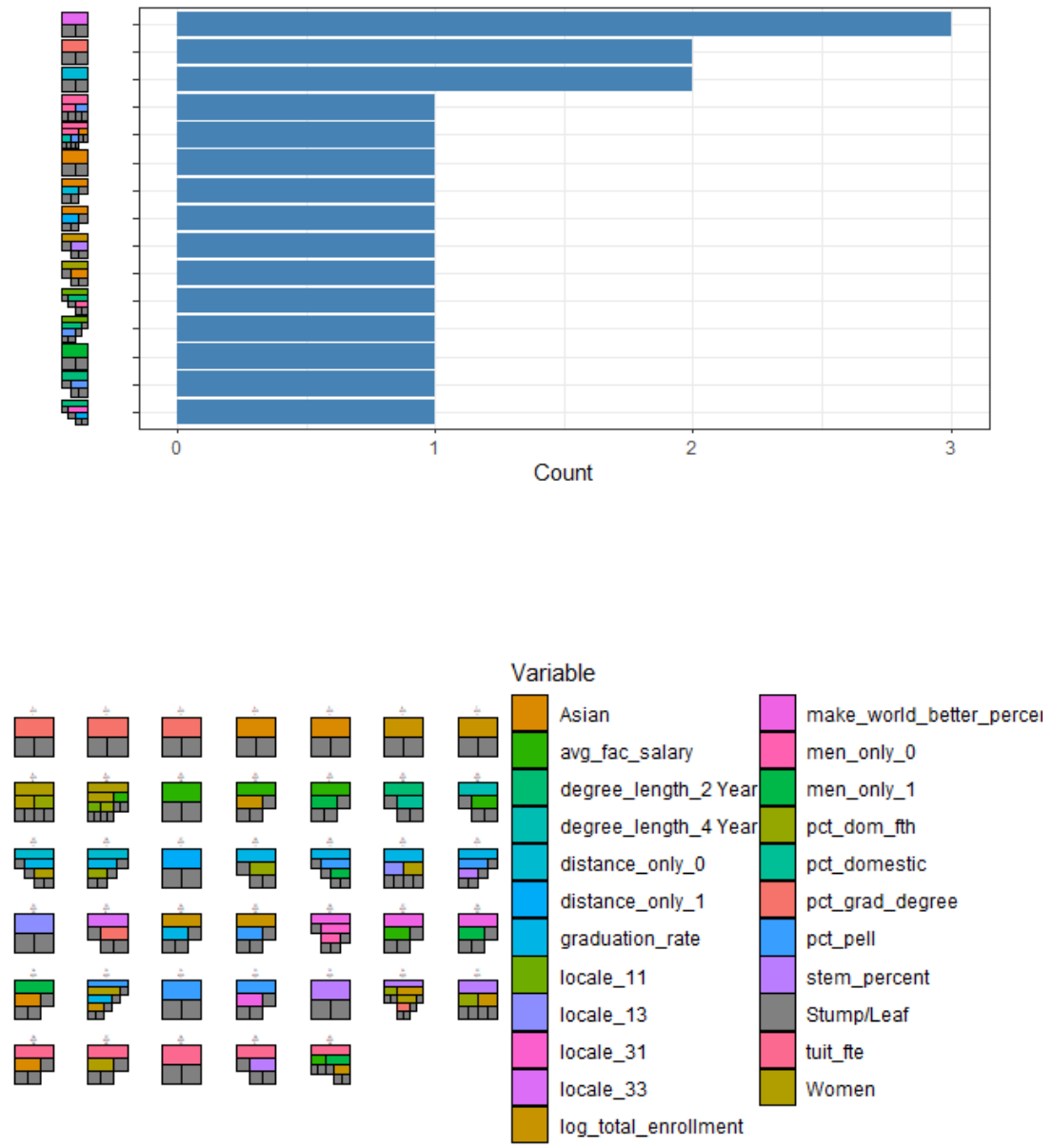

Figure 12: Barplot of 15 most prominent trees in final iteration (Top) with accompanying key of all trees (bottom)