To: Professor A. Diamond

From: William Nguyen

Date: February 26, 2022

Subject: Jobs Skill Training Program for Degree Holders and No-Degree Holders

Executive Summary:

With the lalonde dataset, we are curious to know whether it is better to provide the training program for people with or without high school degrees. I used three main models (multivariate regression, CART and Random Forest) to predict real earnings in 1978 for four types of people, specifically with or without training, and with or without degree. I believe that **people receiving training are predicted to have greater real income than those not attending training.** Furthermore, **the difference in real income between two groups with degrees is predicted to be greater than that between two groups without degrees.** Therefore, **I suggest that the training program should be invested in those with high school degrees and high school should be made more accessible.**

Data Analysis:

	MSE Test Set Error
Regression without Interaction Terms	88442869
Regression with Interaction Terms	90389398
CART without tree pruning	91354591
CART with cross validation for tree pruning	90334424

Random Forest with optimal mtry	86583526

Table 1. The Table reports MSE Test Set Error of different predictive models.

Random Forest seems to minimize testing error the most.

From the table, a basic regression model performs better than one involving interaction terms, as a consequence of overfitting training data. Furthermore, the CART model with cross validation for tree pruning performs slightly better than the CART one without tree pruning. Most importantly, I have chosen the model with least MSE test set error - Random Forest with optimal mtry. Restricting the number of randomly chosen variables at each split helps decorrelate the relationship between predictors. This is a helpful strategy because variables in our study can easily affect each other. To exemplify, race and high school degree relate to each other and can affect chances of unemployment because of external social factors (e.g. race discrimination or lack of educational access). Therefore, Random Forest seems the most potential candidate here.

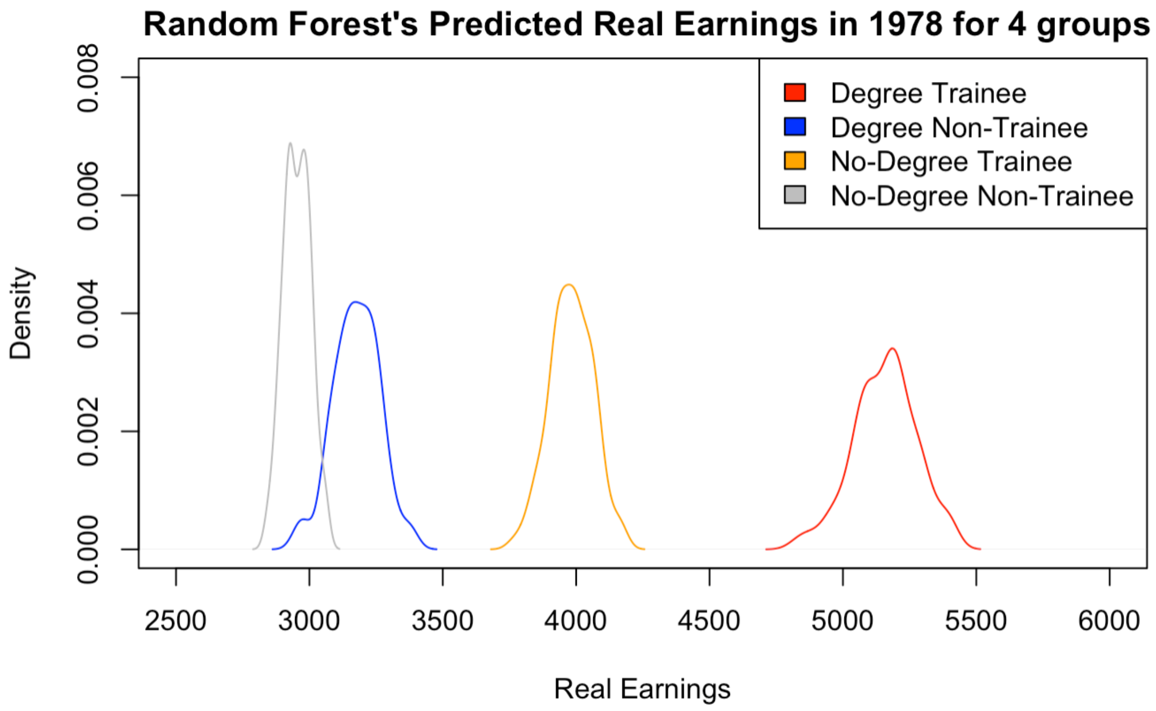

Figure 1. The density visualization incorporates all predicted values for 4 typical subjects from the Random Forest model with the optimal mtry.

I ran the Random Forest model 100 times to add more noise and randomness to the prediction. Then, I plotted 100 predicted values for each category to better visualize and understand the patterns here. It can be clearly seen that without treatment, degree holders are predicted to earn more than those without degrees. However, the interesting part is that with treatment, degree holders are predicted to have a jump in income, from only around 3000 - 3500 to 5000 - 5500. The jump in no degree holders with treatment is predicted with less greatness, from around 3000 to around 4000. Therefore, we can use this predictive inference to decide to invest in more training programs for degree holders.

Now, we examine other models to examine if we can derive similar insights.

Regression:

In multivariate regression, I involved 9 predictors as typical characteristics of a trainee. I have two models, one with interaction terms between having no degrees and receiving the training. According to Table 1, the model without interaction terms performs better. With guesswork, the model with interaction terms learns training data so well that overfitting happens and testing accuracy reduces. Therefore, I simulated my regression model with no interaction terms and added normally distributed noise to the model. After that, I plotted predicted values for 4 typical types.

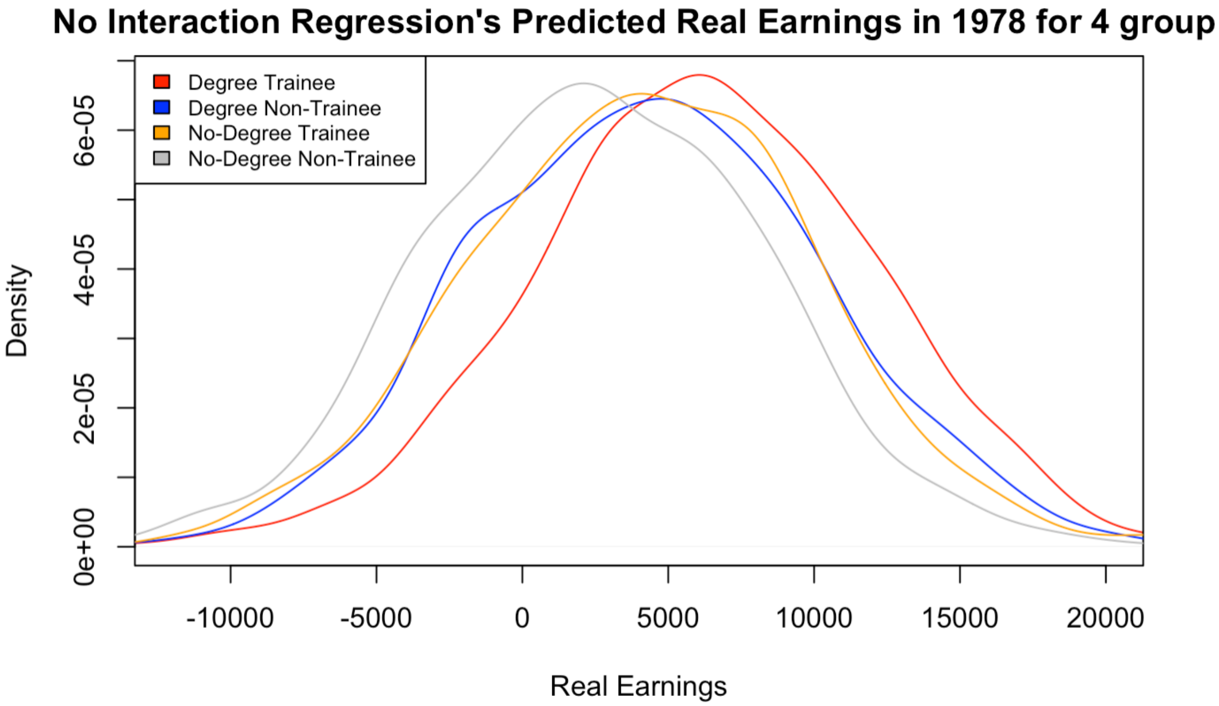

Figure 2. The density visualization compares predicted 1978 real earnings for 4 types of typical subjects in the study.

We can see that subjects having a degree and receiving training are predicted to earn the most out of four categories. Meanwhile, the gray line representing the no degree and no training group is leftmost in the visualization.

CART with cross validation for tree pruning:

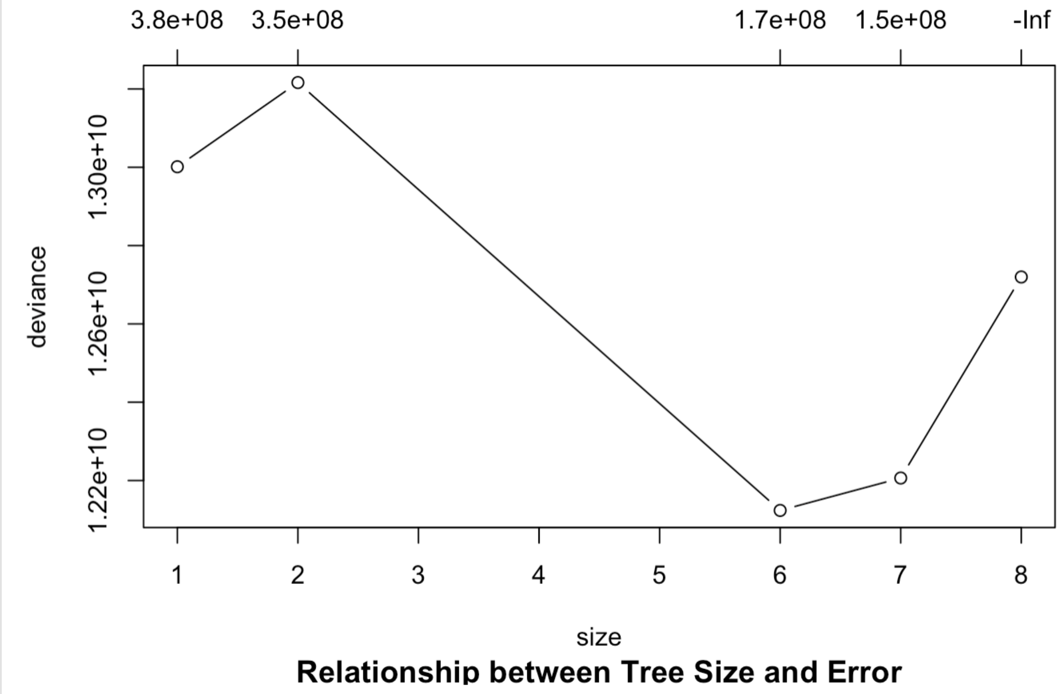

Figure 3. The plot shows the relationship between different tree sizes and their corresponding deviance. From this plot, we can see that 6 is the optimal tree size with smallest error.

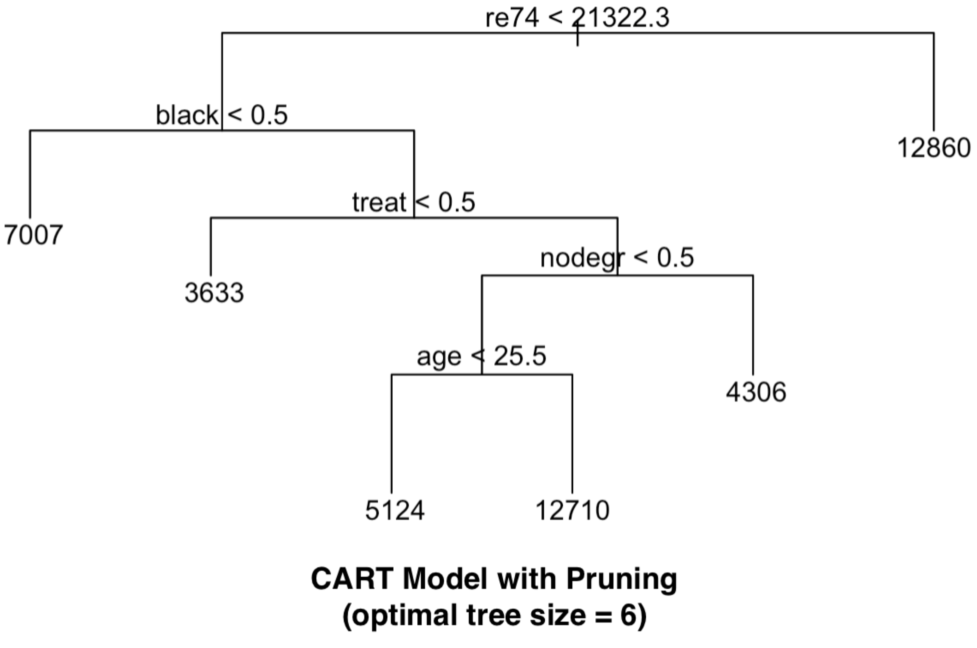

Figure 4. The plot demonstrates our tree model with 5 splits in total. We can see there are 6 terminal nodes.

Degree + Training	Degree + No Training	No Degree + Training	No Degree + No training
5124.077	3633.219	4306.139	3633.219

Table 2. The table summarizes predicted values for 4 groups from the CART model.

With CART, I visualized the relationship between tree size and corresponding deviance, along with the tree model at its optimal size. From Table 2, without training, the typical subject having a degree or no degree is predicted to earn the least, which can be understood from examining the tree (Figure 4). Moreover, with training, people with a degree are predicted to earn more than those without a degree. This result is consistent with my overall findings presented in the Executive Summary.

Random Forest:

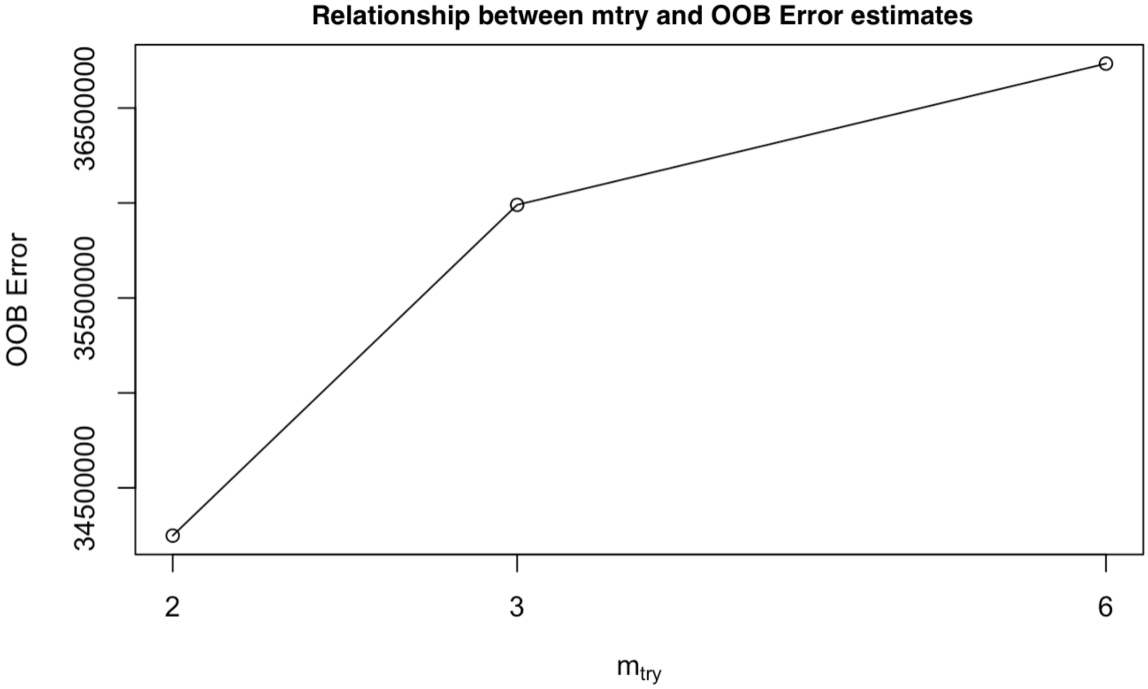

Figure 5. The plot visualizes the number of mtry in the Random Forest model and its corresponding OOB.

mtry	OOB Error
2	34248450
3	35990172
6	36733989

Table 3. The table summarizes OOB Error in each mtry in the Random Forest Model

In the Random Forest model, we can see that with mtry = 2, we can get the smallest OOB Error. We can decorrelate the influence between predictors and add randomization by involving only 2 randomly generated variables in each split.

Limitations:

Categories	Number of Observations	Categories	Number of Observations
Degree	348	Trainees	131
		Not-Trainees	97
No Degree	97	Trainees	54
		Not-Trainees	43

Table 4. The table summarizes the number of observations in different categories in our study

(with / without degrees and with / without training participation)

In our dataset, we see that the number of degree holders is much larger. Therefore, if this study can be replicated with more data points about people without a high school degree, our conclusion can be further supported. Moreover, we have no information about counterfactuals here, which prevents us from reaching causal inference. We can only infer predicted values here for different groups.

Conclusion:

With a predicted pattern of greater improvements in real earnings between groups of degree holders than between groups of no degree holders, **I strongly support more training programs for degree holders**. Furthermore, **high school education should be made available to current no degree holders** so that training's effectiveness can be optimized in this group.

[1]**Word Count**: 744

Appendix:

My code can be found here.

[1] This word count does not involve headings, tables, figures, references or subtitles.