The first step to solving any issue is to examine the available data so that you have a better understanding of what options you have for your model. This data was a mixture of nominal and continuous data, providing for opportunities to condense the data. I started off by performing univariate analysis, looking at the distribution of continuous variables and frequency counts of the nominal data. I looked for outliers and opportunities to condense levels within variables to reduce the variance within our models, or more plainly, to make sure that there was greater consistency and interpretability in the model.

I was concerned with the number of levels in the data, so I examined the relationship between the target variable broken out by each level of the data set. Wherever applicable, I condensed those levels into one, while trying to preserve the information and the distribution. A great example of this would be that there were different levels of education, but the relationship for individuals who never completed high school was approximately the same, providing a chance to reduce dimensionality and increase model effectiveness. I then applied this same methodology for education, workclass, and country; paying close attention to preserving relationships between the levels and the target variable (identifying individuals making $50,000 or more).

This problem of identifying if an individual makes $50,000 or more is a classification issue, meaning that someone is identified as a 1 or 0 in our model, or an individual who makes the target amount or an individual who does not. Given the nature of the data, being both continuous and nominal, I approached this problem with logistic regression and a classification tree. The logistic regression is an equation that provides each level of a variable with a specific odds and then totals those odds together to give us a likelihood that an individual is making greater than $50,000. Based on this model, we can create a cutoff in terms an acceptable likelihood of making $50,000 to determine if an individual is either a 1 or 0. When we draw this line at .50, or an assigned probability of 50% from our model, this regression equation had an accuracy of 83.16%.

As a form of validation, I also created a decision tree which charts out interactions between variables to determine if an individual is making $50,000 or if they are not. This method has the advantage of being very interactable and has answers of yes or not, making it easy to follow along. If an individual meets the conditions identified in the initial node, it splits and goes to the node to the right, following this pattern until it reaches a terminal node identifying if an individual is classified as our target variable. However, when it is fed new information that it was not trained on, it has an accuracy of 82.37% or almost 1% less accuracy than our logistical model. What it makes in interpretability, it suffers regarding prediction accuracy.

Based our model, seen in the attached graphics, there is a strong relationship between education level and the likelihood of an individual making more money. When combined with capital gains, as seen in the jitter plot at the top of the image, we can see that the two together with a very strong indicators of income. Location may provide a influence, as seen in the second graph which shows the proportion of individuals making 50,000, but ultimately developed countries have a greater chance of making higher incomes. Lastly, in the bottom graph, we confirm again the relationships between education and income level, but also adding in the lens of gender, highlighting a discrepancy between males and females in income levels.

