McKayla McAllister

Professor Snehasis Mukhopadhyay

Introduction to Data Science

2 December 2023

1. Problem statement

We must compare different analysis methods to find out which, if any, will pick up on potential biases in the data. Ideally, a machine learning method should be at least 90% accurate to distinguish between whether or not the given person makes more than $50,000 per year. We are testing neural networks and decision trees. We will use the testing data one protected variable at a time (in this case, we will test “male”, then “female”). That way, we have the accuracy for the different protected variables, and we can calculate how different they are from each other to attempt to determine if the bias is too steep for our data. It is important to remember that while methods can be biased, it is also possible that the data that we are using is biased.

1. Methods

As stated above, there were two analysis methods tested.

1. Neural networks

Our neural network will be utilizing a cost equation to determine if it is a good fit or not. We can compare the cost, or a calculated value relating to the difference between the real value and the predicted value, to gauge how off the algorithm usually is. For the previous assignment, I anticipated this method to contain more bias due to the increased number of connections. However, for this assignment, I anticipate the opposite. Due to the increased number of connections between data points, the method will prove to be less biased than decision trees.

1. Decision trees

We can also use decision trees. We can determine if decision trees are inaccurate or biased by comparing their accuracy. In the previous assignment, I anticipated the simpler decision trees to allow for less bias, but it wound up being the opposite. I believe that it is due to the simpler method of taking shortcuts, like gender.

1. Experimental results (combined)
   1. Neural networks

The cost for the combined neural network after 1000 epochs of training was .509. This indicates that the model is very inaccurate. Despite the high cost, the model was not trending any higher for most of the training

* 1. Decision trees

The overall accuracy of the decision tree method was nearly 83%. This indicates that this model is also inaccurate for our results. I wonder if it has something to do with the type of data and the “weight” that each piece of data has (fnl\_wgt). This is a variable with the purpose of indicating how many copies there would be of that row in the full dataset.

1. Experimental results (male and female separated)
   1. Neural networks

After 1000 epochs of training, the cost for male-identifying persons was calculated to be .509 rounded to the nearest thousandth. This, considering that our target variable is binary, is not a great result. As I mentioned with the combined set, the sex-specific sets also appeared to be converging, so additional epochs would likely not have benefitted the model much. With that in mind, I proceeded to change the dataset to test the female-identifying persons, for which the calculated cost was also .509 rounded to the nearest thousandth after 1000 epochs. The cost of the algorithm was .509 for the male set, and .509 for the female set.

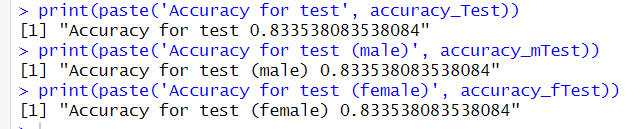
* 1. Decision trees

After proceeding with the decision tree for this dataset, the reported accuracy was only 83%. This would not be enough to be an accurate model for the dataset, although it is better than chance. All of the accuracy tests were extremely close, which indicates that the model did not build any differently between the combined set, the female set, and the male set.

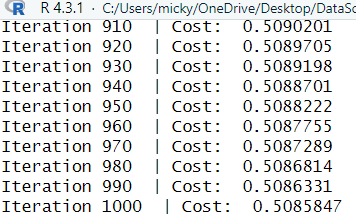
1. Discussion of results

Although both methods passed our test for bias, neither method passed the basic accuracy standard that was set at the beginning of this assignment. This indicates to me that, though the data is not biased in this way, the dataset is not adequate enough for machine learning in these ways. Perhaps integrating the weight variable would have created a dataset that would have been better to learn from. I could have created a method that loops through each entry in the original dataset, and duplicates the row fnl\_wgt number of times. Then, I could have that be my primary dataset, with which I could then divide up based on sex and test like I did with the smaller set. I wonder if that would make a difference, though, or if the data is just skewed somehow else that could be impacting the results. I did try implementing this solution, as I suspect the data is skewed towards uncommon values, but the data frame could not handle the request to duplicate all the columns because of the size.

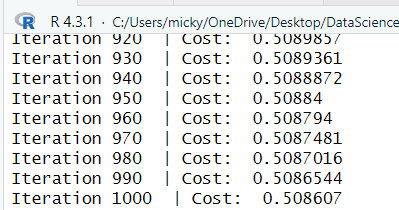
*Image: Decision trees data*

**

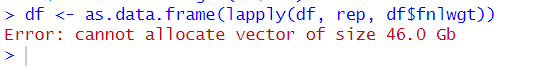
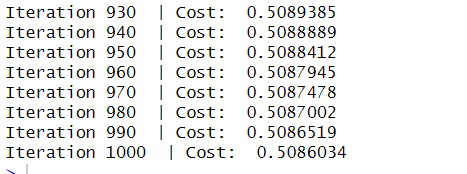
*Image: Neural network data (male)*

**

*Image: Neural network data (female)*

**

*Image: Neural network data (combined)*

**