DSCI 372

Henry Gipson

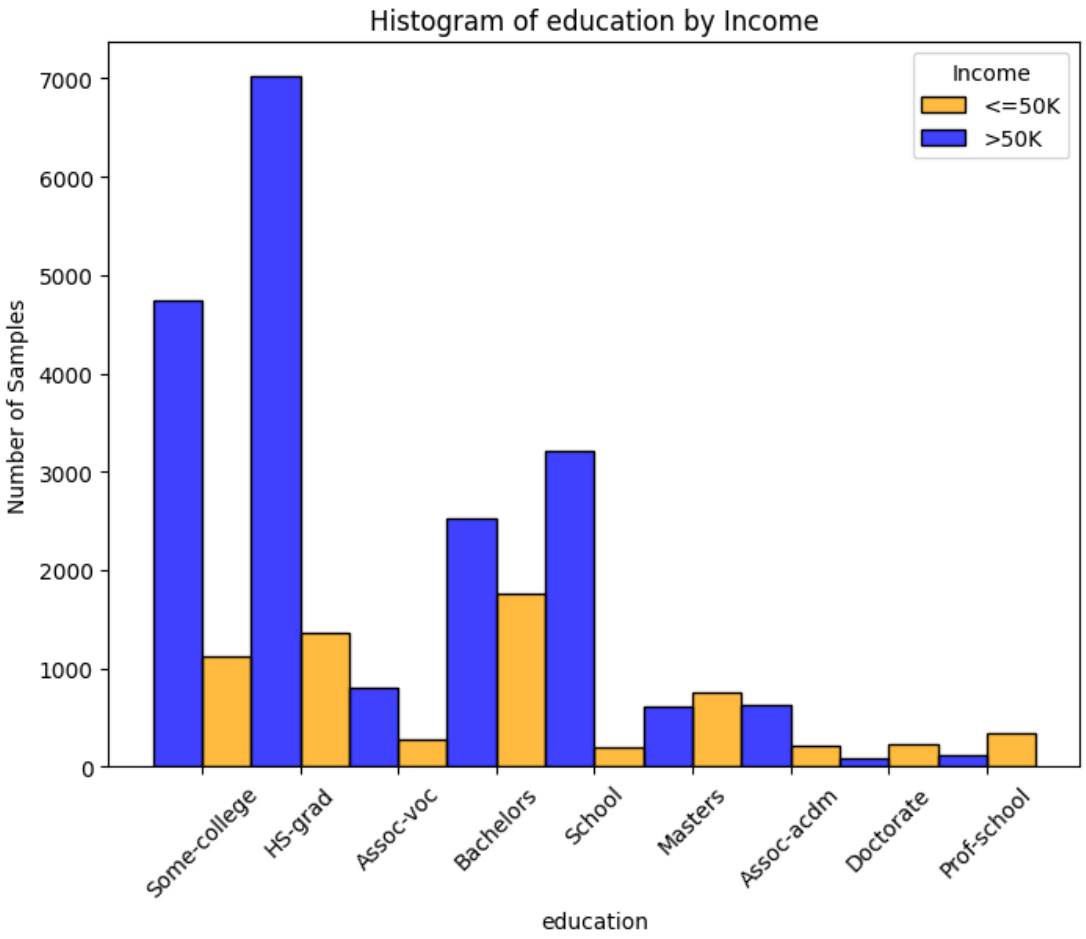
Project 1 Report

1)

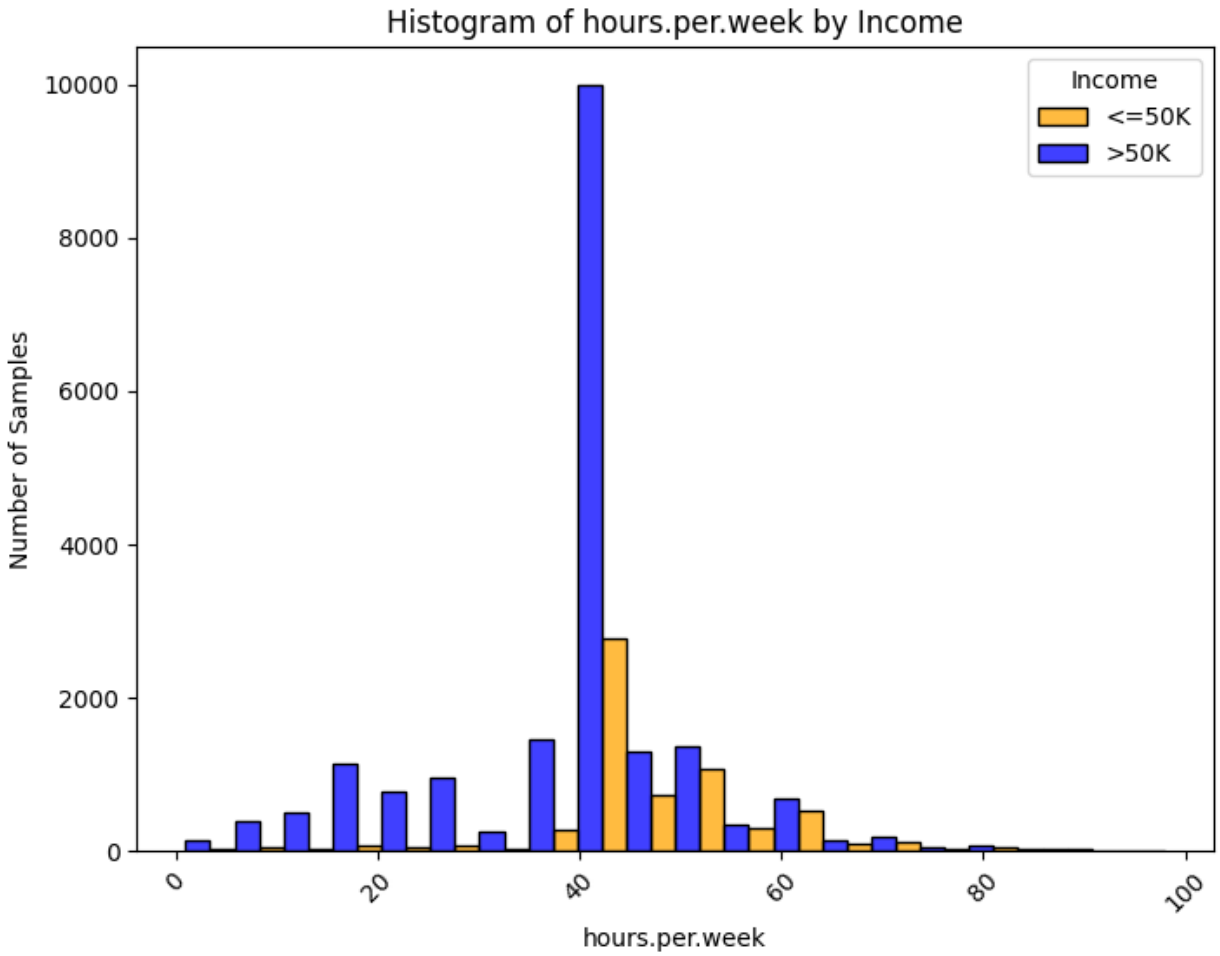
The above plot is the correlation plot of my data analysis. What surprises me here is that capital loss and gain are next to zero in terms of correlation. The feature we are hoping to zone in on the “income” seems to be most correlated with capital gains and least correlated with capital losses, which from a basic understanding of business practices

seems somewhat logical. More gains would suggest more income in a standard capitalistic model.

2)

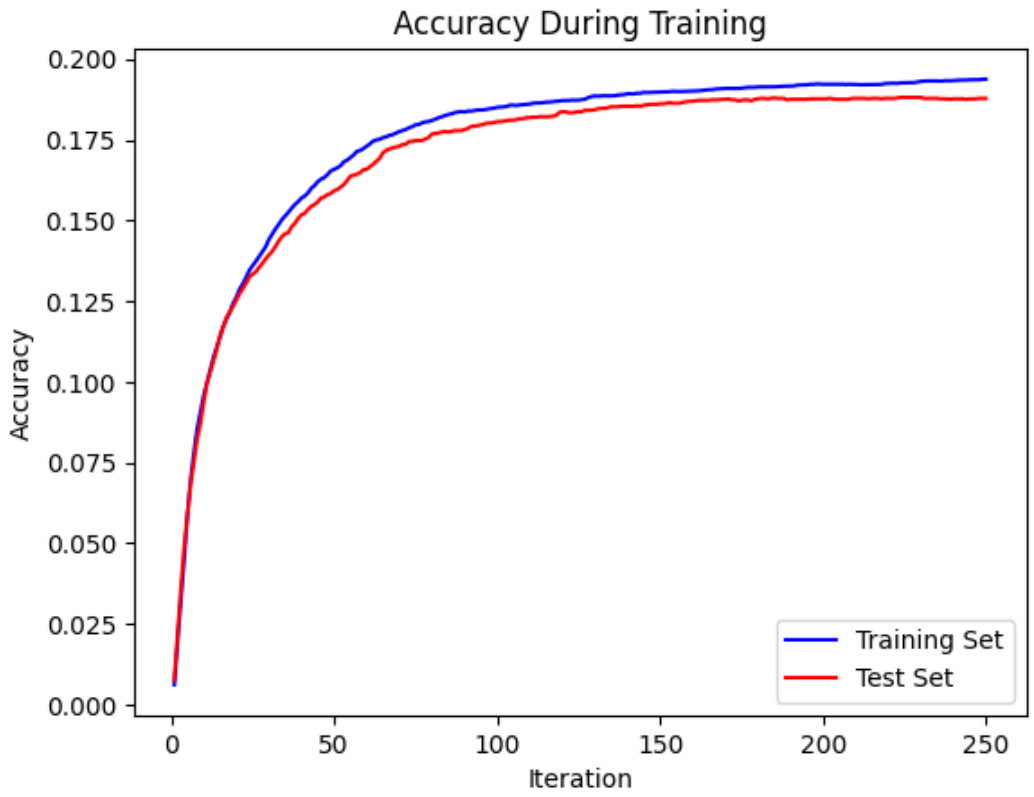


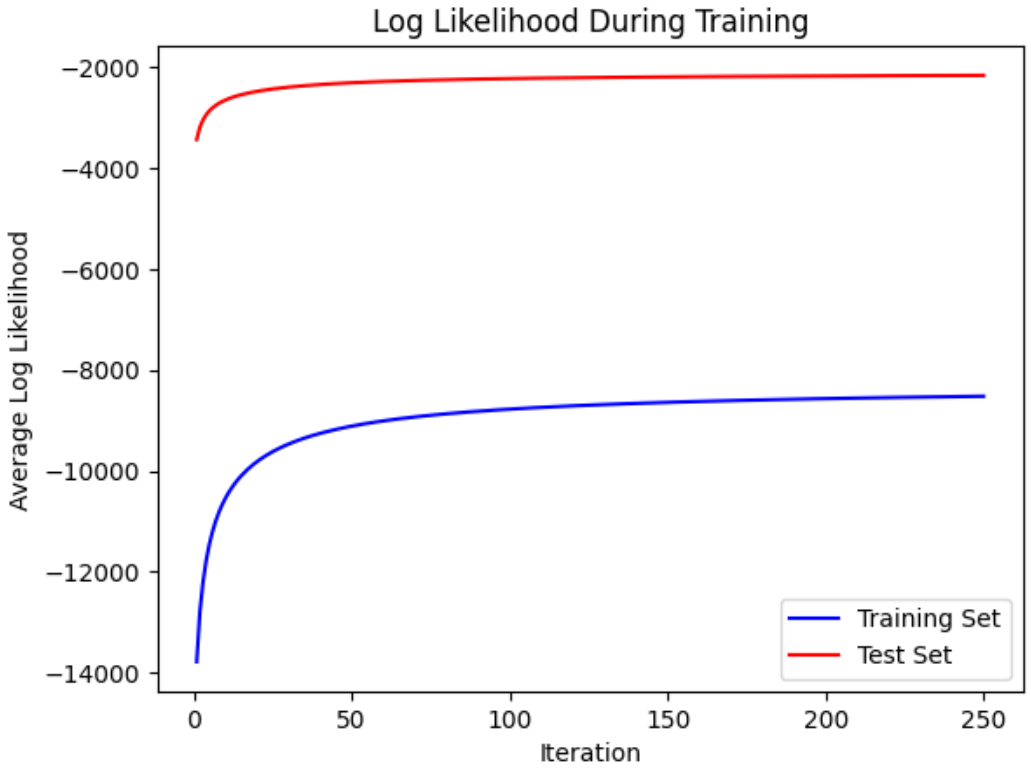
In the above plot we can see that a majority of samples where individuals have an income of over 50K are contained within the HS-Grad category, which I would find to be a surprise. Because of how this plot was generated, this is not caused by some outliers there is simply is that many samples in that category. What questions this brings to mind is the data collection methods, and if that process is air tight then what does this mean about the common standard that college degree lead to more financial success?



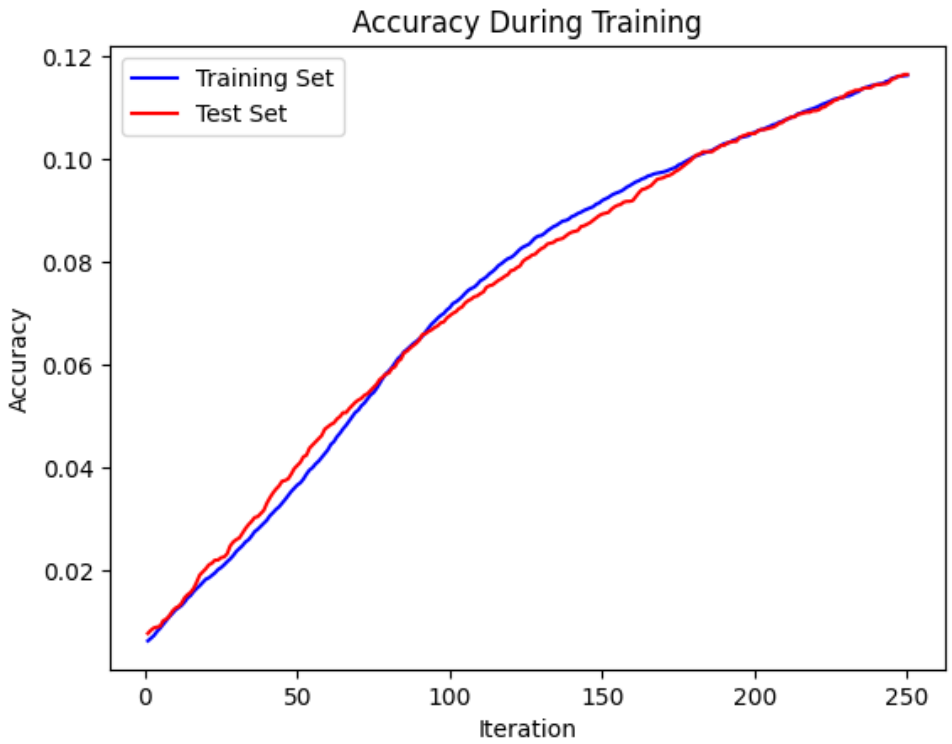
In the plot above here, we can see a clear distinction that the highest density of above 50K is right around 40 hours a week of work. This seems logical as most of the employers in the US currently use a standard 40hr work week, with the federal governmental standard of 40 backing this up. Any hours over 40hs are commonly known as overtime so with this data being taken from a sample of a population it makes sense to see that the average hours of corporate America shows up as the mean.

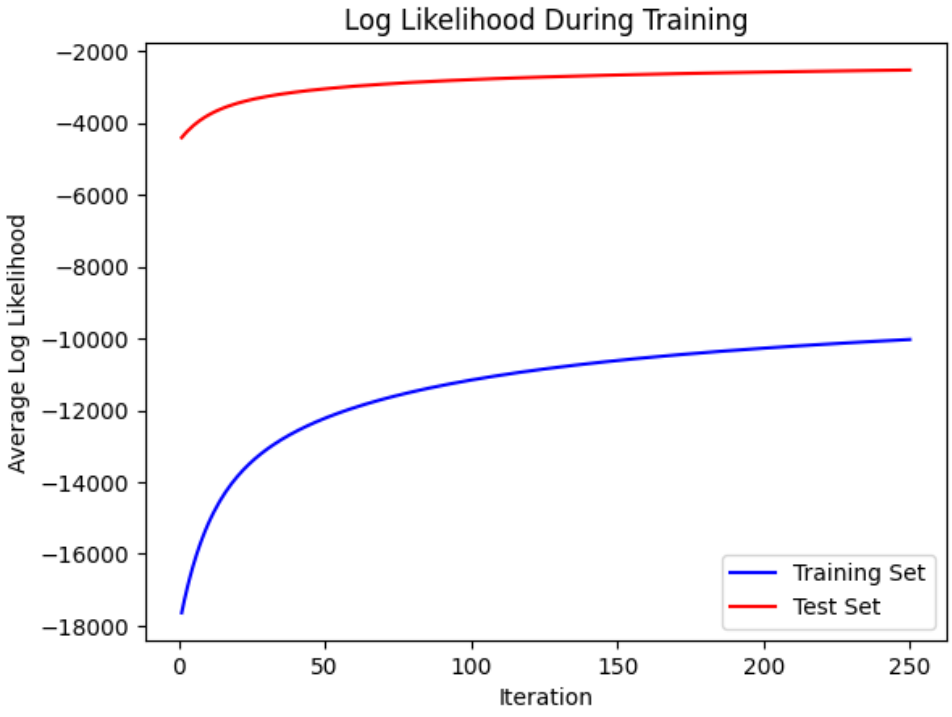
3)





The above plots are for a learning rate of 0.75





The above plots are for a learning rate of 0.5

In the 4 above plots there is a clear distinction, the log likelihood has minimal change between the plots, enough that I had to double-check my code to ensure it was working

properly. The larger learning rate (0.75) saw a drastic spike in accuracy but proceeded to taper off, this is most likely because as it narrowed down on a true value for the weights the “jump was to large and the program proceeded to miss on the sides of the true value.

On the other hand, the smaller learning rate (0.05) continued gradually until the program terminated, this is likely because while it was unable to narrow down quickly the smaller steps with time are more likely to find the true value of the weights.

4)

At the time of submitting this assignment I was unable to complete the ROC figure