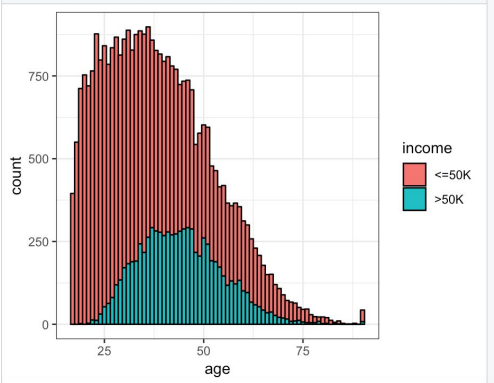
Overview of the Project:

For the Capstone project, I delved into the Adult Census Income Dataset to see what insights I could get and build a prediction model for the data. The Kaggle Dataset detailed incomes for various people around the world. It gave information such as age, work class, education level, marriage status, occupation, relationship status, race, sex, hours worked, and native country, all with a trailing column of whether the person made $50,000 in a year. With nearly 32,000 data points, the dataset was very large and thorough and could have been applied in a variety of different ways. Since the dataset provided if the person made less than or greater than 50k in a year, it was a natural leaping point for creating model and using the model to predict if the person was above or below the value. After cleaning the data, and adjusting the group types, I used logistic regression for the machine learning in the data set.

Methods/Analysis:

The data was at first cleaned and values adjusted, to ensure that the data would be properly parsed and the most important characteristics would properly be represented in the model. I then grouped some of the response types, to ensure that there were not too many points for the model to be parsing, and only logical groups were present for certain data values. For example, the dataset had the options of both “never worked” and “without pay”. Due to the similar nature of both of these values, I grouped them together as “unemployed”. Similarly, in the marriage category, the data could be represented as either “divorced”, “seperated”, or “widowed”. Since they all fall under not being married, I grouped them together, as they all fell under the same category and would not have a large effect in the classification.

After playing around with some machine learning methods, I decided to use logistic regression as it was a logical choice for the data and yielded a really good accuracy rate. After trying both K-means and Random Forest, two choices that I thought could be incorporated well and provide a strong accuracy rate, I went with logistic regression for my choice as it provided the highest rate for accuracy. I first plotted some of the data, to understand the look and get a feel for how the data could be represented. One of the factors I did look at was age, one that I thought would have a big impact on the grouping and classification of the data.



After sifting through the data and the overview of the data, I began to implement the logistic regression for predicting. I first split the data, and the had the data to train the data on. And then train the data with response to the different data points and categories that were available. Since the output I was looking for can either be above or below 50k, I implemented the binary regression to determine which of the two outputs each of the test values would yield.

Results:

After training the data, I tested the data with thousands of data points. Out of the data points, my classification determined that 6,939 of the people would less than 50k and that 477 would make more than 50k. This yielded an accuracy value of 0.8518632, which is equivalent to 85.19%. My prediction used 15 fisher scoring iterations to properly classify the data.

Conclusion:

With the accuracy percentage being so high, the binary logistic regression has made out to have been a good choice for the classification. In the end, the output confusion matrix has turned out to yield fairly accurate results with a residual deviance of 14535 on 22743 degrees of freedom. In conclusion, this project has taught me a lot about real world analysis and classification. The insights I was able to take away from the dataset and the project have not only helped me learn more about the machine learning process but also the world around us.