PART B

1. We were given 11 variables used to decide whether someone would default on their loan or not. With the given data, a loan default would occur approximately 12.5% of the time. Only about half the variables given had a noticeable impact on whether someone would default or not. By far the most glaring was adjusted annual income. The mean annual income for someone who didn’t default was $58,446 compared to someone who did default which was $47,333. This was a difference of over $10,000. The second most notable difference was 58% of people who didn’t default were owning their home while 50.5% of people who did default were renting. People who rented were 8% more likely to default. This makes sense because people who buy are usually a bit more financially stable than those who rent. Lastly, the third noticeable difference was those who defaulted on average had more inquires in the last 6 months at a 20% higher rate. Public record bankruptcies, number accounts ever over a 120-month period, and open accounts made zero difference on whether or not a loan would default. The trends of the rest of the variables matched intuition but you’d see only small differences between the numbers for those who defaulted and those who did not.
2. DATA:

A screenshot of a computer

Description automatically generated

A graph showing a line

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

A graph with a line

Description automatically generated

3.

Both models had an accuracy of between 55 and 60%. I tried dropping variables that looked to have no little effect on a loan defaulting, but my accuracy and ROC changed only slightly. The AUC for logistic regression and naïve bayes was between 60% and 65%. This was better than flat out guessing, but neither model reached the recommended threshold of at least 80%. My guess is this data is missing some other key variables that could benefit the accuracy of the model. Another guess is that the models are too simple compared to the complexity of the algorithm used by the loan companies. They could predict extreme cases, but a fair amount of these loans seem to be right on the edge of defaulting or not. The graph for naïve bayes would grow more sharpy in the beginning stages, while the graph for logistic regression would grow more steadily.