Part A

1. The code creates a logistic model from the customer data using all variables except zip code and their id to predict whether they will take out a personal loan. A logistic model is used to get the probability that an observation can be classified one way or another. If we consider P values under .05 to allow us to reject the null hypothesis, then the output shows that there are eight statistically significant variables: **income**, **family**, **ccavg**, **securities account**, **cd account**, **credit card**, **online**, and the **education level** when *graduate* or *advanced/professional*. Apart from mortgages, it appears that the more products the customer has through the bank the more likely they are to be interested in a personal loan. Therefore, targeting customers that already use many of the banks services may be a good way to proceed.

2. The model would give her a 25.64% probability of being interested in a personal loan. Depending on the selected threshold, if we said 50% and below was a “no” and above 50% was a “yes” then, no she would not be interested. This threshold, according to a quick google search, can best be determined with a ROC curve.

Part B

1. The Naïve Bayes calculates the probability for each variable independently against the target and multiplies it against the other variable probabilities to get a final probability. The KNN model plots each observation and then chooses the closest k outcomes to figure out the target outcome. For KNN I chose a K of the square root of 5000 which was about 71 and turned education into dummy variables since it wasn’t clear if it was a factor variable or not. In retrospect this probably wasn’t necessary since the scale was 1-3 which can be measured between. The decision tree weights each variable to figure out what is most important and then builds a tree of decisions to work through to get to a final classification. I chose a minimum split of one for the decision tree.

2. The output for Naïve Bayes, KNN, and Decision tree gives you a matrix of correct and incorrect classifications against the test data set. Each model can then be used to predict a new observation. In this case the Naïve Bayes and KNN models predicted "not interested" where the Decision Tree predicted "interested". In the case of the Naïve Bayes, Decision Tree, and Logistic Regression you can determine the most significant variables. These significant variables will give you the best **available** predictors to get to your preferred target.

3. The models except for the Decision Tree all predicted that the customer would not have been interested in a personal loan. Although with my current knowledge it is difficult to compare the models directly, these outcomes align to the low probability of being interested in a personal loan that the logistic regression outputted.

4. The decision tree model had the highest predictive accuracy against the test set at 98.3%, with KNN following at 94% accuracy, and the Naïve Bayes getting the lowest with 88.7% accuracy. Although the Decision Tree had the highest accuracy, I have doubts about it because it was the only model that predicted that the customer would have been interested and it seems with an accuracy that high it could have been overfit. I would **not** choose the Naïve Bayes because it assumes independence between variables where I believe there would be a strong correlation between some variables such as using their online banking service and having any single account at the bank. Another example of that variable correlation would be education and income. The logistic model is just difficult to read and reason about. Therefore, from a matter of deduction, I would have chosen the KNN model because it aligned with the other models' predictions and the predictive accuracy was still very high.