**PART 1: Data Exploration and Preprocessing**

1. **Which column represents our target variable?**

The ‘Personal Loan’ column represents our target variable.

1. **How many customers accepted their personal loan offer in the last campaign? How many did not accept their offer?**

480 customers accepted the personal loan offer whereas 637 customers didn’t accept the personal loan offer.

1. **Are there any NA values? If so, which columns are they in?**

Yes, there are a total of 7 null values. 4 null values are in the Experience field and 3 null values are in the Income field.

1. **Which category in the ‘Education’ column occurs most frequently?**

The ‘Undergraduate’ category occurs most frequently (389 times) in the Education column, followed by ‘Advanced’ and ‘Masters’.

1. **Are any of the predictors highly correlated with one another, as defined by a correlation of 0.9 or greater?**

Yes, ‘Age’ and ‘Experience’ are positively correlated with each other with a correlation of 0.99.

1. **Did you drop predictors? Why or why not?**

Yes, we dropped 2 predictors – ‘ID’ and ‘Experience’ from the dataset. ID was a unique identifier for each row and hence would not play any role in predicting the response variable. Experience was found fit to dropped because it was highly correlated with the Age column.

1. **How did you deal with the categorical variable(s)?**

There is only one categorical predictor in the dataset – ‘Education’ This variable was converted into numerical values by using the binary flagging method. After flagging and dropping the first variable, we have 2 new columns that replaced the old Education column –Education\_Masters and Education\_Undergraduate.

1. **Did you scale the data? Why or why not? If you did, what technique did you use?**

Yes, we scaled the data because predictors like ‘Income’ and ‘Mortgage’ have values in thousands unlike other predictors which are in the range of 10s. We used z-score normalization to scale the values of predictors.

**PART 2:**

1. **What is the training and testing performance (F1 score) of the first model where k = 5?**

When k=5,

* Training F1 Score = 94.6
* Test F1 Score = 89.6

1. **How do you think this modeling is performing in its predictive task?**

Based on training and testing performance, the model is doing well in prediction (when k=5), with a F1 score of 89.6. However, it is slightly overfit because the F1 score reduces from 94.5 to 89.6 in the test dataset.

1. **What is the optimal value of k? What is the F1 score of that model on the test data?**

The optimal value of k could be 13. The F1 score is highest for this model = 89.86.

The model is performing well at k = 3, 5, and 7 as well.

**Part 3:**

1. **What is the training and testing performance (F1 score) of the logistic regression model?**

* Training F1 Score = 88.79
* Test F1 Score = 86.11

1. **Is this model overfit?**

Yes, the logistic regression model appears to be slightly overfit because the model is performing better with the training data. The F1 score reduced when the model was used on test data.

1. **Between the logistic regression model and the k-NN model (with the optimal value of k), which model would you recommend for deployment/further testing? Support you answer with sound reasoning.**

We will recommend using the Logistic regression model for deployment or further testing.

If we compare the F1 scores on test data of the two models, we can infer that both the models are almost equally good except that the F1 score of k-NN is a little better. But k-NN is computational heavy as every time a new observation is provided to the model, it must calculate all the distances real time before giving a predicted value. Thus, the prediction phase is slow for large datasets.

Since in this scenario, where we have to classify a person as a loan acceptor or non-acceptor, a Logistic regression model will be more suitable as we will get the results quickly without compromising on the performance significantly.