# Dataset/problem description

Universal Bank is a new bank with a rapidly growing customer base. The majority of these customers are liability customers (depositors with checkings/savings accounts). The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this sector to bring in more loan revenue. In particular, it wants to explore ways of converting its liability customers to personal loan customers. Last year, Universal Bank ran a marketing campaign where they offered personal loans to select customers. The dataset from that campaign is provided in the *UniversalBank unprocessed.csv* file, and the data dictionary is provided below:

1. ID: Customer ID
2. Age: Customer’s age
3. Experience: Number of years of professional experience
4. Income: Annual income of the customer (in thousands USD)
5. Family: Family size of the customer
6. CCAvg: Average spending on credit cards per month (in thousands USD)
7. Education: Education level
8. Mortgage: Value of mortgage held with Universal Bank, if any (in thousands USD)
9. CD Account: Does the customer have a certificate of deposit (CD) account with the bank? (1= yes, 0 = no)
10. Online: Does the customer use internet banking facilities? (1= yes, 0 = no)
11. CreditCard: Does the customer use a credit card issued by Universal Bank? (1= yes, 0 = no)
12. Personal Loan: Did this customer accept the personal loan offered in the last campaign? (1= yes, 0 = no)

Universal Bank has hired you as a data science consultant. Using the given data, you are tasked with building and testing two classifiers (namely, *k*-NN and logistic regression), where the goal is to predict whether or not a customer accepted a personal loan that the bank offered in the last campaign. If you are able to develop a model that is capable of this, then Universal Bank can use the technology to identify which customers are most likely to accept loans in the next marketing campaign. You will compare the predictive performance of the two models, and make a decision regarding which model to “move forward” with.

# Part 1: Data Exploration and Preprocessing

Complete the following tasks within your code:

1. Explore the data following whatever techniques you wish (just be sure you include enough exploratory analysis so that you’re able to answer the questions for the write-up).
2. Drop predictors as you see necessary.
3. Impute any NA values using the *k*-NN imputer (the code for this is provided in the .ipynb file)
4. Deal with categorical variables and scale the data if and how you wish.

Answer the following questions in your write-up:

1. Which column represents our target variable?
2. How many customers accepted their personal loan offer in the last campaign? How many did not accept their offer?
3. Are there any NA values? If so, which columns are they in?
4. Which category in the ‘Education’ column occurs most frequently?
5. Are any of the predictors highly correlated with one another, as defined by a correlation of 0.9 or greater?
6. Did you drop predictors? Why or why not?
7. How did you deal with the categorical variable(s)?
8. Did you scale the data? Why or why not? If you did, what technique did you use?

# Part 2: *k*-NN

Complete the following tasks within your code:

1. Train and test a *k*-NN model with *k*=5
2. Find the optimal value of *k*

Answer the following questions within your write-up:

1. What is the training and testing performance (*F*1 score) of the first model where *k* = 5?
2. How do you think this modeling is performing in its predictive task?
3. What is the optimal value of *k*? What is the *F*1 score of that model on the test data?

# Part 3: Logistic regression and model comparison

Complete the following tasks within your code:

1. Train and test a logistic regression model

Answer the following questions within your write-up:

1. What is the training and testing performance (*F*1 score) of the logistic regression model?
2. Is this model overfit?
3. 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.