Example project: Targeted Direct Marketing

# Business problems and Problem Formulation

Project selection: Targeted Direct Marketing Campaign

## Questions to consider:

1. What is the business problem? Direct marketing, either through mail, email, phone, etc., is a common tactic used to acquire customers. Because resources and a customer's attention is limited, the problem at hand is to increase the number of customers for the banking institution by running a marketing campaign, in this case, a phone campaign.
2. What is the business goal? The business goal is to *only* target the subset of prospects who are likely to engage with a specific offer.
3. What is the business metric? The business metric we will track is the improving number of people who subscribed to a term deposit at the bank.
4. Why is the business problem appropriate for machine learning? This problem is appropriate for ML because there is a lot of data. Predicting potential customers based on readily available information like demographics, past interactions, and environmental factors is common in machine learning. Additionally, patterns for our predictions would be very difficult to code into a set of rules.
5. What type of ML should be used? Binary classification should be used.
6. Reframe the business problem as a machine learning one: Predict if the client will subscribe (yes/no) to a term deposit.

# Iteration I

*Note: In the Student Project Template there is a section for “Iteration II” that includes feature engineering, hyperparameter optimization, and final thoughts. This example is only complete through the first iteration to help get you started.*

# Data preprocessing and visualization

## Questions to consider:

1. Did you have to make any assumptions about the data?
   1. I assumed that the dataset we are using (sample) has the same distribution as the broader population.
2. What does exploratory data analysis and visualization tell you about the data?
   1. After examining the data and applying some visualization techniques, we could see that almost 90% of the values for our target variable y are "no", so we can see that most customers did not subscribe to a term deposit.
   2. We also saw that many of the predictive features take on values of "unknown". Some are more common than others. We should think carefully as to what causes a value of "unknown" (are these customers non-representative in some way?) and how that data should be handled.
      1. Even if "unknown" is included as its own distinct category, what does it mean given that, in reality, those observations likely fall within one of the other categories of that feature?
   3. Many of the predictive features have categories with very few observations in them. If we find a small category to be highly predictive of our target outcome, do we have enough evidence to make a generalization about that?
   4. Contact timing is particularly skewed. We noticed that almost a third of the people were contacted in May and less than 1% in December. What does this mean for predicting our target variable next December?
   5. There are no missing values in our numeric features. Or, missing values have already been imputed.
      1. pdays takes a value near 1000 for almost all customers. Likely a placeholder value signifying no previous contact.
   6. Several numeric features have a very long tail. Do we need to handle these few observations with extremely large values differently?
   7. Several numeric features (particularly the macroeconomic ones) occur in distinct buckets. Should these be treated as categorical?
   8. Customers who are in the "blue-collar", "married", "unknown" default status, and were contacted by "telephone", and/or in "may" are a substantially lower portion of "yes" than "no" for subscribing.
   9. Distributions for numeric variables are different across "yes" and "no" subscribing groups, but the relationships may not be straightforward or obvious.
3. What techniques did you use to clean and preprocess your data?
   1. We handled missing values

# Model training

## Questions to consider:

1. What percentage of the data should be training, validation, and test?
   1. We randomly sorted the data then split out first 70%, second 20%, and then 10%
2. Did you randomize the split? If not, how might that impact the model?
   1. Yes. We randomly split the data.
3. What algorithm should be used?
   1. We are using XGBoost but we can use Linear learner as well.

# Model evaluation

## Questions to consider:

1. What metric did you choose to evaluate your model?
   1. We used accuracy to determine how many people subscribed to the term deposit.
2. How did your model do on your chosen metric?
   1. Of the approximately 4,000 potential customers, we predicted 136 would subscribe and 94 of them actually did. We also had 389 subscribers who subscribed that we did not predict would.
3. What did you learn from your evaluation metric?
   1. We learned that in its current state this model is producing less than desirable results, but it can (and should) be tuned to improve.