DSO 510 Assignment 1

Problem Statement

The goal is to help the bank do the solicitation campaign specifically to those who would most possibly respond. And possible data resources are:

(1) Behavior data, as well as PII, of 10 million current customers, including 100,000 who apply for an installment loan(referred as loan thereafter);
(2) A few million records of PII data about those participated in previous solicitation campaigns, a couple hundred thousands of which are of those who responded;
(3) Demographic and psychographic data, as well as PII, about prospects (bought from data broker).
(4) Demographic and psychographic data about current customers and those participated in previous solicitation campaigns (bought from data broker).

We can make a summary table as below.

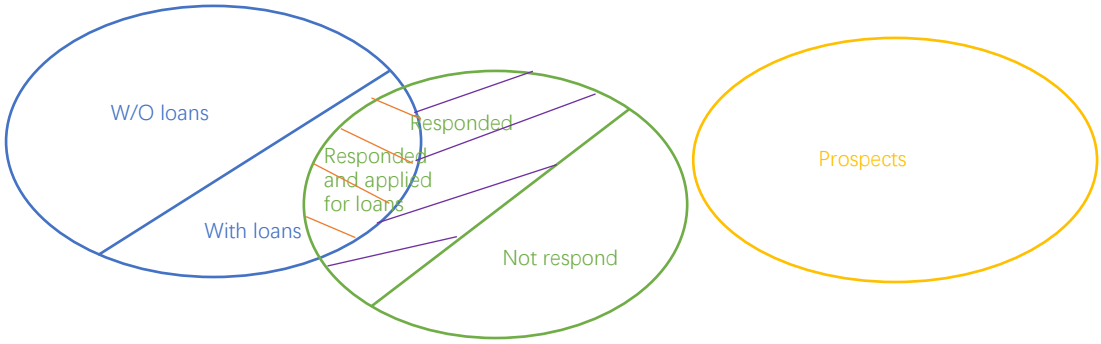
Table Summary of Data Resources

Cluster	Mini-cluster	# of records	Data Type
Current customers	With loans	10 million	PII, behavior data, demographic and psychographic data
	W/O loans	100,000	
Previous campaigns	Responded	A few million	PII, demographic and psychographic data
	Not respond	A couple hundred thousands	
Prospects	/	infinite	PII, demographic and psychographic data

As we can see, there's a difference between data about current customers and data about prospects. Consequently, the behavior data we have on hand is kind of useless since we have no access to these behavior data of prospects. Even though we could use behavior data as inputs to do modeling, we could not use such data to make predictions on prospects. In this case, behavior data is completely useless.

However, if luckily, our behavior data on hand has some kind of demographic and psychographic data within it and these overlapped data are sufficient to do modeling with high scores(e.g. accuracy, AUC), then these behavior data would be much helpful. Our following proposal is made on this assumption.

Most-likely-respond prospects



Graph Venn diagram of Clusters (regardless of cluster size)

Take a closer look at clusters, we can see that people responded in previous campaigns are separated into two groups:

- Group A (the area of red lines): those who responded as well as applying for loans.
- Group B (the area of purple lines): those who responded yet did not apply for loans.

Therefore, when targeting at most-likely-respond prospects, we should split the target group into two corresponding groups:

- Group A: those who are already interested in applying for an installment loan and then solicited to pick our bank as their choice.
  - About Group A, our data on hand is not only their PII, but also their behavior data since they have applied for loans and therefore became our current customers.
- Group B: those who were purely solicited by the campaign without urgent needs for loans.
  - About Group B, our data on hand is only their PII.

Proposed Analytics Framework

Separate Group A from Group B

Now that we have different source of data about two different groups of people, we should split the problem into two different models using different data. So the very first thing to do is to separate Group A from Group B.

- We could use our customers as proxies for those who responded. Specifically, customers with loans represents people who responded and applied for loans, while customers without loans represents people who responded yet did not apply for loans.
  - With our on-hand data, we could create a supervised model to predict whether or not responded people would apply for loans. In other words, we will find out people with existing needs for loans and therefore are much more likely to be solicited by and respond to our

campaign.

Targeting at Group A

Once we make the model, we could then let the data broker run our model to find out those classified as interested in loans (i.e. Group A) and then launch the solicitation campaign towards them.

Targeting at Group B

For Group B, since we only have their PII, we definitely have to buy data from brokers in order to do modeling. Below is my detailed proposal:

- Step 1: Find out PII about previous campaigned people and then use PII to particularly buy their demographic and psychographic data from data brokers.
  - We should at least buy 2,000 records each for those responded and those didn't.
- Step 2: Create a supervised model, with response as the binary output, to predict prospects' response to the solicitation campaign.
  - Since we buy data from brokers, we could use people's demographic and psychographic data as inputs, whose dimension should be large enough for modeling.
  - Make sure we buy more records than the dimension of one record to avoid ill-posed problems.
- Step 3: Let the data broker run our model to find out those classified as people who respond(i.e. Group B), and then launch the solicitation campaign towards them.

Further clarification on the proposal

There're certain points I want to elaborate on.

- Separating Group A from Group B, rather than treat them as one and use response as the only output to do modeling, provides three advantages:
  - Save money spending on buying data.
    - Our model inputs to find out Group A does not need data from data brokers.
  - More precise predictions are made based on two models compared to one model.
    - The magic of modeling is to find common things behind data. Though there must be common features for those who respond to the campaign, they could be driven by completely different motivations. By separating two different groups, we could find out different features working behind different motivations and therefore make more precise predictions.
  - The Group A model can be further exploited when targeting at maximizing applied loans.
- If my assumption, that behavior data on hand has some kind of

demographic data within it and these overlapped data are sufficient to do modeling with high scores, is too hard to achieve, then we could use the proposal specifically for Group B to do the entire modeling and buy more records than total 4,000, to ensure high scores of model.