DSO 510 Assignment 2

Problem Statement

The goal is to help ATT build segment labels using their CDR records, which can be further sold to other companies as a valid variable for their models.

Model Design

Model

There are several segmentation methods, like clustering, decision tree, and Hierarchical Clustering Algorithms (HCA), to adopt for our goal. Because currently I don't have any available segment labels, I would use clustering to do the unsupervised learning.

Input

Since ATT only has CDR info and PII, I'd like to buy some other data like demographic and/or psychographic data (referred as external data afterwards) to help build the model.

In order to make segment labels universal to all ATT clients, I have to let the data broker run my model to output segment labels, since I could not afford to buy external data for all my clients. In this way, due to legal issue, I have to use summary CDR info as inputs of the model.

- Detailed info about specific individuals would be lost if I use summary CDR info, which is harmful for model accuracy. In order to make my labels more precise on an individual basis, I would use summary CDR info based on geographical region zip9, trying to minimize the information loss.
- On the other hand, if it is legal to send detailed CDR info to the data broker to run the model (I am not familiar with the legal issues), I would definitely adopt detailed CDR info as inputs of the model.

A sample input would be consisted by:

PII(phone-number level), CDR(summarized), external data(bought according to PII).

Output

I'd like to create about 35 segments, and make detailed descriptions for each segment. And I would do this on a phone-number level in case anyone has more than one phone number.

Proposed Framework

- First I have to summarize my CDR info. I would average CDR info on the zip9 level and then replace original CDR info with averaged CDR info according to their zip9.
- Then I would randomly pick 8,000 phone numbers as my observations. Through PII, I could buy their demographic and/or psychographic data from

data brokers.
    - I'm picking 8,000 phone numbers rather than all phone numbers due to the data cost for buying their external data.
- After gathering all the inputs needed, I would build the unsupervised clustering model to create 35 segments.
- Once the model is built, I would send my summarized CDR info of rest phone numbers, along with the model to let the data broker run my model and then get segment labels for all my phone numbers.
- I would do explorative analysis, on a segment level, to come up with detailed descriptions for each segment.