# Homework 15

## Question 20.1

*Given the following data sets:*

| ***Data set #1*** <br> ***(from an alumni magazine publisher)*** | ***Data set #2*** <br> ***(from a credit bureau)*** | ***Data set #3*** <br> ***(collected by the company using web site tracking code)*** |
|---|---|---|
| *first name* | *first name* | *title* |
| *last name* | — | *first name* |
| *college or university attended* | *middle name* | *middle initial* |
| *year of graduation* | *last name* | *last name* |
| *major or majors* | *marital status* | *credit card type* |
| *marital status* | *sex* | *credit card number* |
| *number of children* | *year of birth* | *list of products purchased in the past, with date of purchase and ship-to address* |
| *current city* | *current city* | *which web pages the person looked at* |
| *email domain* | *whether they ever owned real estate* | *how long the person spent on each page* |
| *financial net worth* | *email domain* | *what the person clicked on each page* |
| *binary variables for interests* | *list of monthly payment status over the last five years for credit cards, mortgages, rent, utility bills, etc.* | *estimate of how long the user's eyes spent on each page viewed* |

*How could the company use these data sets to generate value, and what analytics models might they need to do it?*

### Solution

In my opinion, an initial classification approach as opposed to just clustering clients would yield more useful results since we could add a scientific reasoning into what we are classifying. For instance, with a clustering approach, we could find a general Fitness related cluster because people with hobbies that are related to sport will probably buy fitness products or supplements. Even though everyone in that category is related in some products, their main goal could be different. Moreover, there are many people that practice sports for the sole purpose of keeping a healthy mental state or to avoid "stress". Therefore, the goal of that person is not fitness, and perhaps we ought to provide him products such as sleeping pills (clonazepam so he can get addicted and buy more) or melatonin, rather than advertising some random zinc supplements just because he is in the fitness cluster.

**Step 1: Classify the customer into a correct, pre – established category.**

**Given:**

- Customer ID (pre-matched customer data across all data sets with name data, credit data, birth date, sex, marital state, email domain)
- Age
- Marital status

- Data about interests (from data set #1)
- list of products purchased in the past, with date of purchase and ship-to address, which web pages the person looked at, how long the person spent on each page and what the person clicked on that page (from data set #3)

**Use:** k-nearest neighbours or support vector machine with supervised learning.

**To:** Classify customer into a category of products. For instance, a customer could be classified in the "Supplements" category. Sub categories could also exist as well, for instance we could have a "Fitness supplements" or "Sleep supplements", etc.

Secondly, we need to predict the budget of our customer. Let's say for instance that our customer is indeed into gaming and in the page looks for a gaming mouse and keyboard and other related products. His budget will be of interest since we could predict if we need to advertise "premium" products, or cheap products. For instance, if he stares longer at cheaper products, or if doesn't get a good economical status predicted from credit data, we ought to advertise cheaper products. Imagine that the customer has a low budget and is hesitant to buy a new keyboard. If we show him a "Razer" (a premium gaming accessories brand) overprized product, he might close the page immediately. On the other hand, if we offer him some "Red Dragon" product (another gaming accessories brand), he might get tempted and look further.

For the following Model, we use financial data as predictors and the web page (which product he tends to look more) as the results.

**Step 2: Predict customer budget.**

**Given:**

- Customer ID
- Type of credit card, monthly payment status, whether they ever owned real estate, Financial net worth, Student status
- Which web pages the person looked at, how long the person spent on each page, what the person clicked on each page
- Results: List of products purchased in the past, with date of purchase and ship-to address

**Use:** Logistic regression.

**To:** Determine the probability that the customer will buy a premium or non-premium product given a certain budget.

Finally, we can design experiments to improve our advertisement testing different product and different versions of the ad using step 1 and 2.

**Step 3: Design of experiment**

**Given:**

- Customer ID
- Classified customers from Step 1.
- Different version of ads for different classification of products taking into account budget prediction from Step 2.

**Use:** Multi – Armed Bandits.

**To:** Determine which products or versions of product ads to show next

The idea is to draw a random sample from a pool of similarly correlated product. By that, I mean that we shouldn't only consider that the products look similar (by image recognition). Even though this might be useful in the case of clothes, as explained earlier, it's not as effective if we are advertising supplements. Finally, if the person clicks on the add, move to exploitation, perhaps he will be willing to expend a few extra pounds for a premium quality product that really satisfies what he's looking for.