Improving Marketing Efficiency Using Predictive Analytics for M&S (I): Customer Segmentation Using K-Means Clustering*

MSIN0094 Case Study

Dr Wei Miao

October 23, 2024

1 Predictive Analytics and Marketing

Machine learning (ML) refers to the study of methods or algorithms designed to learn the underlying patterns in the data and make predictions based on these patterns. A key characteristic of predictive analytics techniques is their ability to produce accurate out-of-sample predictions. Consider the problem of predicting whether a user will click on an ad. We do not have a comprehensive theory of users' clicking behavior. Predictive analytics methods can automatically learn which of these factors affect user behavior and how they interact with each other, potentially in a highly non-linear fashion, to derive the best functional form that explains user behavior virtually in real time. Predictive analytics methods typically assume a model or structure to learn, but they use a general class of models that can be very rich.

Predictive analytics models can be divided into two major groups: supervised learning and unsupervised learning.

Supervised learning requires input data that has both predictor (independent) variables and a target (dependent) variable whose value is to be estimated. If the goal of an analysis is to predict the value of some target variable (e.g., whether customer responds to our marketing offers; whether customers churn at some point in time), then supervised learning is used.

On the other hand, **unsupervised learning** does not identify a target variable, but rather treats all of the variables equally as inputs. In this case, the goal is not to predict the value of a variable, but rather to look for patterns, groupings, or other ways to characterize the data that may lead to an understanding of the way the data interrelate. Clustering analysis is an example of unsupervised learning, which helps data analysts find customer segments based on provided characteristics.

In this case study, we are going to analyze the same dataset as in "Descriptive Analytics for M&S". Our task is to use predictive analytics tools to help M&S conduct more effective targeted marketing.

As a quick recap, the variable definitions are as follows,

Demographic Variables

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company

Customer Purchase History Data

- ID: Customer's unique identifier
- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- NumDealsPurchases: Number of purchases made with a discount
- NumWebPurchases: Number of purchases made through the company's web site
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's web site in the last month
- Complain: 1 if customer complained in the last 2 years, 0 otherwise
- **Response: 1 if customer accepted the offer in the last campaign, 0 otherwise**
- Recency: Number of days since customer's last purchase

2 Tom's Return from the Maldives: A Predictive Analytics Adventure

After a much-needed holiday in the Maldives, where he spent his days sipping bubble teas (yes, even in the Maldives there are bubble tea shops), Tom finally returned to face reality. Feeling recharged, he opened his laptop only to be greeted by the ominous sight of the M&S dataset waiting for him. His task? To clean up the data before diving into predictive analytics for customer segmentation and targeting.

"I'm a proud graduate of UCL's BA program," Tom thought confidently. "How hard can this be?"

He launched his Jupyter notebook, cracked his knuckles, and started writing Pandas code. But after just a few lines, reality hit him hard–his data wrangling looked like a tangled mess of spaghetti!

Without the beloved pipe operators in Python, Tom was forced to write a new line of code for each data wrangling step. Before long, his intermediate datasets were scattered all over the place, and the operations were so unintuitive that he began questioning why anyone would willingly subject themselves to Python for data wrangling. It also brought back some, well, traumatizing memories of the weekly Python quizzes he had to endure during his time at UCL.

"Why did I even bother with those dreadful Python quizzes when R is the best language in the world?" Tom muttered under his breath. "I really should have listened to Dr. Meow, who kept encouraging us to practice R during those Thursday Python classes."

Tom tried to refocus and remember what he'd learned about using dplyr for data wrangling back in Week 3 of the Marketing Analytics class. But unfortunately, that particular week he was a little distracted—-too busy swiping through Tinder in search of a Halloween date rather than paying attention to Dr. Meow's lecture. "If only I had focused on the magic of dplyr instead of debating whether to go as Harry Potter or a giant bubble tea for Halloween!" Tom sighed.

Realizing he was in trouble, Tom did what any smart person would do in this situation: he turned to you, a rising star in data science from this year's BA program. "Please," he pleaded, "help me clean this M&S dataset using R. I need to show M&S how to segment and target their customers more effectively with predictive analytics, and I can't afford another disaster."

And so, your task begins: to help Tom make sense of the M&S data using the power of dplyr and predictive analytics techniques. In return, Tom promised that the next time he goes on holiday, he'll bring you back a giant bubble tea from the Maldives.

2.1 Data Collection

Please load the M&S dataset from your local machine. You can use the "import dataset" button in RStudio to upload the dataset. The dataset is named `data_full.csv`. Name the dataset `data_full`.

> **Question 1**
> load data_full, create `total_spending`, and select `total_spending` and `Income` as the clustering variables into a new data frame `data_kmeans`.

2.2 Data Preprocessing

First, we need to check missing values and resolve them as the k-means algorithm cannot handle missing values directly.

> **Question 2**
> We find that the `Income` variable has missing values. We can replace the missing values with the mean of the `Income` variable.

Next, we need to re-scale the two variables using `scale()`, because the two variables are of very different scales.

> **Question 3**
> Scale the variables and create a new data frame `data_kmeans_scaled`.

2.3 Data Analytics

Question 4

- Apply K-Means Clustering with 2 Clusters

Question 5

- We can examine the structure of the `result_kmeans` using `tidy()`

Question 6

- Visualize the clustering

Question 7

- Determine the optimal number of clusters using statistical criteria

2.4 Business Recommendations

After segmenting the customers into two groups, we can now analyze the two segments to understand their characteristics and behaviors For example, we can compare the average response rates to marketing offers in the two segments, and decide which segment to serve when launching the next marketing compagine to save marketing costs.

As we would like to save marketing costs by sending marketing offers to responsive customers, we need to compute the average response rate for each segment generated by the K-means algorithm

- Generate the customer segment in `data_full`
- Compute the average response rate for each segment

We observe that Segment 1 has a much higher average response rate of more than 20%, so we should send marketing offers to Segment 1 customers.