WWW Challenge

Briefing

My company, Wondeful Wines of the World (WWW), has the mission of providing its customers with excellent wines. Wines are sold via catalogue, website and 10 small shops. At the same time, and as a company is essentially for wine connoisseurs, people for whom wine is a passion and a hobby, we also sell accessories related to the theme, such as corkscrews, wine racks, tasting glasses, etc.

I recently managed to acquire an excellent model of corkscrew extractors, which I intend to sell to my customer base. Marketing will be based on a mailing to be sent to customers that I think may be interested in purchasing it. In order to be able to define which customers I should send the purchase proposal to, avoiding sending it to everyone and the corresponding costs, I carried out a test. From my database, where I record all the information about my 10,000 clients, I took a sample. This sample was collected at random and is made up of 2 000 customers. For these 2 000 selected customers I then sent the proposal to purchase the corkscrew. After a month of waiting, I already have the results, that is, I already know which of those who chose to buy the corkscrew. Based on this file (2000 contacts and their answers) I need you to tell me which of the 8000 I should contact.

What I need is for you to tell me which customers, out of the 8000, are most likely to buy the corkscrew. You have at your disposal the file with the 2000 customers with a set of characterization variables, including a binary variable that indicates whether the customer joined the promotion or not (Spcork - 0 did not buy, 1 bought). In addition, you also have the 8000 customers on which they must comment, that is, say which ones should receive the proposal to acquire the wine cellar.

Further, I also need to know what the different groups of customers are that I have, based on how valuable they are to my company. I want to understand the different characteristics of each group so that I can develop a relevant marketing plan for each one. Creating this value-based segmentation should lead you to select a subset of variables that are relevant to this task.

Deliverables

1. A csv file with 8001 rows and 2 columns:
   - **Custid** : Customer ID number
   - **Prediction** : 1 if the customer should receive the proposal, 0 otherwise

   The file should be named using this format:

   `FAI2223_Group_99_Prediction.csv`

   where "99" should be your group number. An example csv submission file will be provided on Moodle.

2. A Colab notebook (ipynb file format) containing all the code used to develop your models.

   The file should be named using this format:

   `FAI2223_Group_99_Notebook.ipynb`

   where "99" should be your group number. An example notebook submission file will be provided on Moodle.

3. A written report describing:
   - The results and conclusions you gained from the analysis you performed.
   - The different clusters you found.
   - The process by which you arrived at these insights.
   - What kind of preprocessing you performed, if any, and what algorithms you used. No need to discuss **how** the algorithms work.

   The file should be named using this format:

   `FAI2223_Group_99_Report.pdf`

   where "99" should be your group number.

   Maximum 10 pages of content (excluding cover page, index and appendices).

Available Data

Excel file with the 2000 customers contacted in the test phase and their response, and 8000 customers not yet contacted.

Variables in the database

Name	Values	Statistics	Meaning
CUSTID	1001-10000		customer ID number
DAYSWUS	550-1250	mean=899	number of days as a customer
AGE	18-78	mean=48	customer's age or imputed age
EDUC	12-20	mean=16.7	years of education
INCOME	$10K-$140K	mean=$70K	household income
FREQ	1-56	mean=15	number of purchases
RECENCY	0-550	mean=62	# days since last purchase
MONETARY	$6-$3052	mean=$623	total sales to this person
Spcork	0, 1	7,25%	1=bought the cork Extractor