Data Science Assignment

Objective
This exercise has two parts.

1. The first part is meant to explore creative problem solving
2. The second part is to demonstrate data processing and statistical skills

Part One: Creative Problem Solving
Consider a scenario where we have search data pertaining to brands, more specifically, people who have searched for a brand like Nike. We also have comprehensive browsing and app usage data for these individuals.

Due to data collection cycles there is a delay from data collection to having sufficient data for reporting. Since data is collected historically, it can take time to build sufficient sample that can be used to represent search of a particular brand. This limitation is due to data being collected retroactively.

For example, in month one, we have search data for 1,000 individuals that covered 6 months previous. In month two, we have search data for 1,000 additional individuals. Thus the prior month now has data for 2,000 individuals, and so on.

The end result is that we are only comfortable providing data to customers after three months.

Challenge: How would/could/should we approach this to be able to provide some data signal in less than three months? Are there any methods or techniques that should be considered?

Part Two: Data processing and Statistical skills

We have provided three data files in CSV.

1. Prime_day_purchases_2024.csv
2. Amazon_user_behaviors.csv
3. Amazon_non_users_behavior.csv

Datasets #1 and #2 include purchases leading up to Prime day in October plus browsing behaviors for these users. This creates a comprehensive look at user behaviors and Amazon purchases.

Dataset #3 includes browsing data for users who did not make a purchase during the same time.

What we would like you to explore is can we make predictions whether a user is likely to make a purchase at all. How would you approach this problem and what techniques?

Please use these files to do some data exploration and attempt to build a purchase prediction model.

Deliverables

1. A simple writeup of your approach
2. Any scripting (python or other) to accomplish this
3. An explanation of whether this is a valid approach and what are its shortcomings (or benefits)

Please include any relevant statistical analysis including any descriptive and/or inferential statistics.