Pathsetter
LEARN. GROW. THRIVE

Case-study : Dim Reduction Market Seg

Pathsetter folks have so far witnessed cases and ML algorithms that fall under the category of the supervised ML algorithms. This means the data directly provided some structure and clarity on what the independent variable and the dependent variables are and thus indirectly what needs to be predicted based on what. Now imagine a situation where you are just given a data set with a large number of columns and you are asked to find out “meaningful patterns” _in the data. Feels like being left in a forest (not a “random forest”) blindfolded right? So, what can be done about this 1) The dimensions (no. of columns) are huge 2) There are no clear target and explanatory variables. This is exactly where unsupervised learning technique comes to your rescue. It helps you identify patterns in data sets containing data points that are neither classified nor labeled.

In this case, we have consumer purchase data from an e-commerce giant in the U.S that has both brick-and-mortar stores as well as online services. The company has run several ad campaigns throughout the year targeting its customer base. Different campaigns were effective in converting customers with characteristic features. They need a good marketing strategy now. A marketing strategy is a business game plan for reaching prospective consumers and generating sales. Placing the right product and messaging in front of the right prospect requires understanding customer characteristics and behaviors. Customer insights drive the creation and deployment of effective campaigns with the highest customer response percentage, leading to increased sales. The goal is to understand the problem by analyzing existing marketing data to identify critical insights around customer characteristics and habits, spending and purchase patterns, and then segment customers into groups with similar traits. But before even getting into the problem of which campaign was effective for whom, the dimension of the data set is humongous. Something must be done on that front.

Jeff, the CEO of the e-commerce giant is visiting Hyderabad for the inauguration of its India headquarters and our Srinivasan (Srini) from Pathsetter has found a way to briefly connect with Jeff regarding placement opportunities for Pathsetters’s students. Jeff being Jeff, hands over this following data set and tells Srini in an American accent “This data is too big in dimension for any meaningful analysis, pls ask your student to impress me with a dimensionality reduction technique but……without losing any important information. we can then talk about placements and internships”. Finally, Srini provides the attaches .csv file and the below data dictionary to you seeking meaningful solutions. Can you use the skills that you will learn this week to help Jeff with his dimensionality issue and in turn help Srini/Pathsetter bag some nice placement opportunities?

The data description:

Marketing section of Jeff’s firm provided a dataset from the year 2016. The dataset contains the following features:

1. ID: Unique ID of each customer
2. Year_Birth: Customer’s year of birth
3. Education: Customer's level of education
4. Marital_Status: Customer's marital status
5. Kidhome: Number of small children in customer's household
6. Teenhome: Number of teenagers in customer's household
7. Income: Customer's yearly household income in USD
8. Recency: Number of days since the last purchase
9. Dt_Customer: Date of customer's enrolment with the company
10. MntFishProducts: The amount spent on fish products in the last 2 years
11. MntMeatProducts: The amount spent on meat products in the last 2 years
12. MntFruits: The amount spent on fruits products in the last 2 years
13. MntSweetProducts: Amount spent on sweet products in the last 2 years
14. MntWines: The amount spent on wine products in the last 2 years
15. MntGoldProds: The amount spent on gold products in the last 2 years
16. NumDealsPurchases: Number of purchases made with discount
17. NumCatalogePurchases: Number of purchases made using a catalogue (buying goods to be shipped through the mail)
18. NumStorePurchases: Number of purchases made directly in stores
19. NumWebPurchases: Number of purchases made through the company's website
20. NumWebVisitsMonth: Number of visits to the company's website in the last month
21. AcceptedCmp1: 1 if customer accepted the offer in the first campaign, 0 otherwise
22. AcceptedCmp2: 1 if customer accepted the offer in the second campaign, 0 otherwise
23. AcceptedCmp3: 1 if customer accepted the offer in the third campaign, 0 otherwise
24. AcceptedCmp4: 1 if customer accepted the offer in the fourth campaign, 0 otherwise
25. AcceptedCmp5: 1 if customer accepted the offer in the fifth campaign, 0 otherwise
26. Response: 1 if customer accepted the offer in the last campaign, 0 otherwise
27. Complain: 1 If the customer complained in the last 2 years, 0 otherwise

Assignment 1: Please do an EDA on the dataset and submit this by 24th Feb. Focus on the following aspects during the EDA.

1. How will you address null values and what is your strategy for addressing the outliers in the data?
2. What are the business insights on univariate analysis of customer earnings, customer spending, and website vs store purchases among customers?
3. How is the income level of customers associated with % acceptance across different campaigns? Do you find any other interesting correlation between variables?

Assignment 2: Rerun the analysis that the instructor covered in class and answer to the following questions.

1. In simple English, explain the difference between PCA and t-SNE
2. How many Principal components are required to retain at least 70% of the variance (info) in the data?
3. Take the first 3 principal component in your analysis. What actual variables in the data are the most associated with the 3 principal components?
4. Can you identify any meaningful pattern in the t-SNE analysis?

Note: All the answers and relevant code must be submitted in a Jupyter notebook and descriptive answers must be written in markdown cells. Pls, Submit them before 27th Feb,2023.