Where champions are made
Sportify

Introduction

Welcome to **Sportify**! As the premier destination for top-quality sports products and gear, we are committed to excellence and fueled by a passion for active living. Our curated selection of products is tailored to meet the diverse needs of athletes and enthusiasts alike.

In the constantly evolving landscape of modern business, understanding our customers is paramount to achieving sustained success and fostering enduring partnerships. To this end, we conducted a meticulous examination of three essential datasets, each providing valuable insights into distinct aspects of our customer base.

The first dataset serves as a comprehensive repository of digital interactions, capturing detailed engagements across various online platforms, from email interactions to social media engagement and in-app activities.

Complementing this digital perspective, the second dataset delves into sports product consumption patterns, analyzing spending across a range of activities.

The third dataset focuses on demographic information, providing insights into customer characteristics such as age, city, dependents, and education level.

By applying clustering techniques to these datasets, we transform information into actionable segments, revealing discrete customer profiles to guide our strategic initiatives. Identifying hidden clusters within our customer base enables us to develop fitted marketing strategies and product offerings that align with the unique preferences of each segment.

Project Goal

As a new group of data scientists within our company, you are tasked with developing innovative strategies and insights leveraging our vast repository of customer data. This includes identifying key trends, uncovering hidden patterns, and segmenting our customer base to optimize marketing efforts and enhance the overall customer experience.

The datasets

Digital_Contact.csv

Attribute	Description
Customer_ID	Customer’s unique identification
Email_Clicks	The number of times a customer has clicked on email advertisements or links sent
App_Clicks	The frequency with which a customer interacts with the company's mobile application, such as clicking on product listings, accessing features, or making purchases
SM_Comments	The number of comments made by a customer on the company's social media posts
SM_Likes	The number of times a customer has liked posts or content shared
SM_Shares	The frequency with which a customer shares content or posts from the company
SM_Clicks	The number of times a customer has clicked on links or advertisements shared by the company on social media platforms

What does the company say about this dataset?

Marketing Department: “In this dataset we capture diverse customer engagements, from email and app interactions to social media activities. Analyzing clicks, likes, shares, and comments, we aim to gain valuable insights to personalize marketing and improve customer experiences.”

Data Science Department: “For our department, this dataset presents a unique opportunity to apply clustering algorithms to uncover meaningful customer segments based on their engagement behaviors. By analyzing attributes such as email clicks, app interactions, and social media engagements, we aim to identify distinct groups of customers with similar patterns.”

Data Quality Assurance Department: "We are confident in the integrity of this dataset, as it undergoes regular updates with customer purchase data. However, we believe that we could have some problems with missing values in this dataset.”

Business Stakeholders: Using this specific dataset, we aim to answer the following question:

"What distinct customer segments exist based on their engagement behavior across email, mobile app, and social media platforms?"

Products.xlsx

Attribute	Description
Customer_ID	Customer's unique identification
Fitness&Gym	Money spent on Fitness & Gym Products
OutdoorActivities	Money spent on Outdoor activities products
TeamGames	Money spent on Team Games products
Hiking&Running	Money spent on Hiking and Running products
Last_Purchase	Date from the last purchase
TotalProducts	Number of products bought by the customer

What does the company say about this dataset?

Marketing Department: "In this dataset, we capture valuable insights into customer purchasing behaviors across various product categories, including fitness & gym, outdoor activities, team games, and hiking & running. Understanding spending patterns allows us to tailor marketing strategies and product recommendations to meet the diverse needs of our customers."

Data Science Department: "This dataset provides an opportunity for our department to apply advanced analytics techniques to segment customers based on their purchasing behaviors. By analyzing attributes such as spending on different product categories and the frequency of purchases, we aim to identify distinct customer segments. This segmentation will enable us to personalize marketing efforts and optimize product offerings."

Data Quality Assurance Department: "We believe that this dataset is error-free, since the data available is automatically updated with customer purchases."

Business Stakeholders: Using this dataset, we seek to address the following question:

"Can we identify unique customer segments by analyzing how they purchase across various product categories and their buying patterns?"

Demographic.txt

Attribute	Description
Customer_ID	Customer’s unique identification
name	Customer’s Name
birth_year	Customer’s Year of Birth
City	Customer’s City
dependents	If the Customer has dependents (0 – No; 1 - Yes)
education_level	Customer’s Education Level

What does the company say about this dataset?

Marketing Department: "This demographic dataset enhances our understanding of customer segments by providing insights into age, city, dependents, and education level. Integrating this data with clustering analysis results, we can tailor marketing strategies to specific demographic groups."

Data Science Department: "The demographic dataset enriches clustering analysis with additional attributes."

Data Quality Assurance Department: "We believe that this dataset may contain errors and problems. Thorough preprocessing will ensure data integrity in the demographic dataset, minimizing errors, and enhancing the reliability of insights for strategic decision-making."

Business Stakeholders: "Integrating demographic data refines segmentation insights. We aim to better understand whether the clusters predefined in the previous datasets have distinct socio-demographic characteristics among them."

Deliverables and Evaluation

1. A Jupiter notebook with all the needed code implemented to obtain the results presented in the report and to obtain the results explored in the report.

   The file naming format should be ”DSML202324_Cluster_GroupXX_Notebook.ipynb”, where ”GroupXX” should be your group number.

2. A report that describes the analytical processes and the conclusions obtained. A project that focuses only on the techniques and methodologies approached during the practical classes will have at most 16 values. The remaining 4 values are possible to achieve if contributions based on self-study and creativity are applied, and clearly explained in the report.

   a. It should contain the following structure (text in figures and tables do not count as words):

CHAPTER	MAXIMUM WORDS	MAXIMUM POINTS
Abstract	250	0.5 (2.5%)
1. Exploration	4000	3 (15%)
2. Preprocessing		3 (15%)
3. Modelling		1.5 (7.5%)
4. Description of Resulting Clusters		4 (20%)
5. Action Plan		1 (5%)
6. Conclusion		1 (5%)
References	NA	2 (10%)
Report Quality and StoryTelling	NA	
Creativity and Other Self-Study (Optional)		
Annex1 – KNN Imputer	400	0.5 (2.5%)
Annex 2 – Silhouette Method	400	0.5 (2.5%)
Annex 3 – Other Clustering Algorithm	1000	1 (5%)
Annex 4 – PCA	1000	1 (5%)
Annex 5 – Others	1000	1 (5%)
TOTAL		20 (100%)

b. The font formatting should respect the following conventions:
   - Heading 1: Arial, Size 12 pt, in bold
   - Heading 2 (if needed): Arial, Size 11 pt, in bold and italic
   - Text: Arial, Size 11 pt, line space of 1.5 points.
   - Margins: The default ones in word (Top, Bottom, Left and Right as 1”).

c. The reports that do not follow the specified conditions will suffer penalization on the grade.

d. The file naming format should be ”DSML202324_Cluster_GroupXX_Report.pdf”, where ”GroupXX” should be your group number.

e. All chapters and moments of evaluation are graded through a comparison of the work provided by the different groups.

Notes

1. The deadline for the project is the 28th of April 2024. You may submit your project (notebook and report) up to three days after the deadline, until the 1st of May 2024. However, for each day of delay, a deduction of one point (out of 20) will be applied to the final evaluation score.
2. We will evaluate all the topics mentioned based on the report - a well-structured and succinct report will have a big weight on the evaluation.
3. The jupyter notebook will be analyzed only if some doubt arises during the report evaluation. If some steps were done in the Jupyter notebook but not described in the report, we will not evaluate those. As an example, imagine you check the outliers, and at the end of your project, you decide to keep them. In the report, you should mention how you check if you had outliers, what steps were taken to remove them and why you decide to keep them in the end, among other insights that can be relevant.
4. The jupyter notebook should be delivered with all the cells already run and the outputs visible.
5. The report and the code will pass through a process of plagiarism checking.
6. Theoretical context about algorithms / techniques applied should only be provided when those approaches were not given during the practical classes.

Chapter Description

Abstract: A small summary of your work. The abstract should give an overview of your work: What is the context? What is your main hypothesis? What did you do? What were your main results and what main conclusions did you draw from them.

Data Exploration: Describe the data available and extract meaningful insights that may be helpful in addressing the problem at hand.

Preprocessing: This stage includes all the steps from raw data into data ready for clustering: data cleaning, transformation, and reduction (when needed). It also entails business-related transformations of the input features, creating new features if feasible, and accompanying explanations. Here, we evaluated the quality of your implementation, its justifications, and the insights extracted from this stage.

Modelling: Implementation of K-Means algorithm on dataset “Digital_Contact” and on “Products”. You should, at least, consider those two perspectives. More perspectives are optional and considered as points in the “Creativity and other selfstudy - Others” section. The number of clusters should be defined using the elbow method.

Description of Resulting Clusters: Each cluster should be statistically and visually explored and described, emphasizing the characteristics that differentiate them. A final analysis on the resulting clusters provided by the concatenation of the clusters from the previous chapter “Modelling” (on Digital_Contact and on Products) with demographic data should be explored in this chapter.

Action Plan: You should provide a succinct but well-oriented action plan that includes recommendations on how Sportify can leverage the insights obtained from your analysis.

Conclusions: The critical ideas discussed throughout the project should be summarized and emphasized.

ANNEXES: Creativity and Other Self-Studies (Optional)

If other techniques not covered during practical classes are applied, a theoretical explanation of the algorithm/technique should be provided. In these topics, it is expected that students explore other concepts and techniques not covered during the theoretical and/or practical classes, not only in the theoretical context and implementation but also in the analysis of the outputs and possible comparison of those with the default ones applied in the remaining chapters.

As an example (but not exclusively):

- KNN Imputer should be compared with more traditional approaches to fill missing values. This comparison may include, as an example, changes in the descriptive summaries of the variables.
- The Silhouette method should be compared with the elbow method when defining the right number of clusters. This comparison can include, as an example, possible changes in the resulting clusters.
- Other clustering algorithms should be compared with the results of clustering obtained using K-Means, with further interpretation and analysis of those results.
- PCA should be compared with the results of clustering obtained using K-Means using the individual variables, with further interpretation and analysis of those results.

Good work!