Graded Project

Machine Learning - Unsupervised Learning

Domain:

- E-commerce

Business Context:

- Customer segmentation is one of the most important marketing tools at your disposal, because it can help a business to better understand its target audience. This is because it groups customers based on common characteristics.
- Segmentation can be based on the customer's habits and lifestyle, in particular, their buying habits. Different age groups, for example, tend to spend their money in different ways, so brands need to be aware of who exactly is buying their product.
- Segmentation also focuses more on the personality of the consumer, including their opinions, interests, reviews, and rating. Breaking down a large customer base into more manageable clusters, making it easier to identify your target audience and launch campaigns and promote the business to the most relevant people

Dataset Description:

The dataset contains measurements of clothing fit from RentTheRunway. RentTheRunWay is a unique platform that allows women to rent clothes for various occasions. The collected data is of several categories. This dataset contains self-reported fit feedback from customers as well as other side information like reviews, ratings, product categories, catalog sizes, customers' measurements (etc.)

Attribute Information:

SL.No	Attribute	Description
1.	user_id	a unique id for the customer
2.	item_id	unique product id
3.	weight	weight measurement of customer
4.	rented for	purpose clothing was rented for
5.	body type	body type of customer
6.	review_text	review given by the customer
7.	size	the standardized size of the product
8.	rating	rating for the product
9.	age	age of the customer
10.	category	the category of the product
11.	bust size	bust measurement of customer
12.	height	height of the customer
13.	review_date	date when the review was written
14.	fit	fit feedback

Data Citation:

- Rishabh Misra, Mengting Wan, Julian McAuley "Decomposing Fit Semantics for Product Size Recommendation in Metric Spaces". RecSys, 2018.
- Rishabh Misra, Jigyasa Grover "Sculpting Data for ML: The first act of Machine Learning". 2021.

Project Objective:

Based on the given users and items data of an e-commerce company, segment the similar user and items into suitable clusters. Analyze the clusters and provide your insights to help the organization promote their business.

Steps to the project: [Total score: 50 points]

- Import the required libraries and load the data: [Score: 3 point]
  1. Load the required libraries and read the dataset. (1)
  2. Check the first few samples, shape, info of the data and try to familiarize yourself with different features (2)
- Data cleansing and Exploratory data analysis: [Score: 20 point]
  3. Check if there are any duplicate records in the dataset? If any, drop them.(1)
  4. Drop the columns which you think redundant for the analysis.(Hint: drop columns like ‘id’, ‘review’) (1)
  5. Check the column 'weight', Is there any presence of string data? If yes, remove the string data and convert to float. (Hint: 'weight' has the suffix as lbs) (2)
  6. Check the unique categories for the column 'rented for' and group 'party: cocktail' category with 'party'. (2)
  7. The column 'height' is in feet with a quotation mark, Convert to inches with float datatype. (3)
  8. Check for missing values in each column of the dataset? If it exists, impute them with appropriate methods. (3)
  9. Check the statistical summary for the numerical and categorical columns and write your findings. (3)
  10. Are there outliers present in the column age? If yes, treat them with the appropriate method. (3)
  11. Check the distribution of the different categories in the column 'rented for' using appropriate plot. (2)

- Data Preparation for model building: [Score: 2 point]

12. Encode the categorical variables in the dataset. (1)

13. Standardize the data, so that the values are within a particular range. (1)

- Principal Component Analysis and Clustering: [Score: 23 point]

14. Apply PCA on the above dataset and determine the number of PCA components to be used so that 90-95% of the variance in data is explained by the same. (7)

15. Apply K-means clustering and segment the data. (You may use original data or PCA transformed data) (8)

    a. Find the optimal K Value using elbow plot for K Means clustering.

    b. Build a Kmeans clustering model using the obtained optimal K value from the elbow plot.

    c. Compute silhouette score for evaluating the quality of the K Means clustering technique.

16. Apply Agglomerative clustering and segment the data. (You may use original data or PCA transformed data) (8)

    a. Find the optimal K Value using dendrogram for Agglomerative clustering.

    b. Build a Agglomerative clustering model using the obtained optimal K value observed from dendrogram.

    c. Compute silhouette score for evaluating the quality of the Agglomerative clustering technique. (Hint: Take a sample of the dataset for agglomerative clustering to reduce the computational time)

- Conclusion : [Score: 2 point]

17. Perform cluster analysis by doing bivariate analysis between cluster labels and different features and write your conclusion on the results. (2)

Submission:

- Please submit the solution file in .html and .ipynb format on Olympus
- Add necessary comments wherever required.