H&M Personalized Fashion Recommendation

CSCI-B 565 Fall 2022

Prithviraj Maurya
Data Science
Indiana University
pmaurya@iu.edu

Shambhavi Shukla
Computer Science
Indiana University
shuklsh@iu.edu

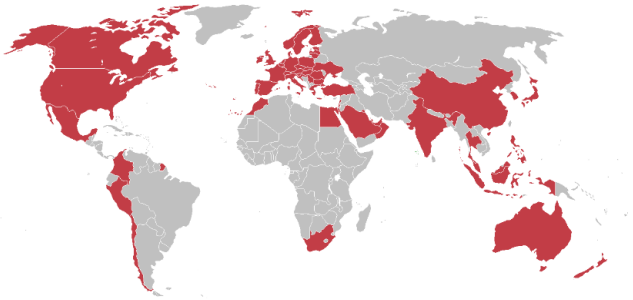
Rajesh Sharma
Computer Science
Indiana University
rajeshar@iu.edu

***Abstract*— The H&M Group is a collection of brands and companies with about 4,850 physical stores and 53 online marketplaces. Online retailers provide customers with a wide range of products to browse. With too many alternatives, buyers might not immediately locate what they want or what intrigues them, which might discourage them from making a purchase. To enhance the purchasing experience, product recommendations are crucial. More significantly, helping customers make wise decisions improves sustainability because doing so reduces costs and, consequently, emissions associated with transportation. Based on information from historical transactions, as well as from customer and product meta data, we aim to offer our customers personalized fashion recommendations tailored specifically to their preferences.**

***Keywords*— Exploratory Data Analysis, Recommendation System, Clustering, Apriori algorithm.**

I. Introduction

H&M is a very popular Swedish clothing company with headquarters in Stockholm. The company was founded in 1947 and was originally named Hennes. In 1968, they acquired the hunting from this point, it operated under Hennes & Maurtiz or simply H&M. In 2000, they entered the US market. As of 2022 it operates shops in 74 countries:



Our work aims to provide H&M's customers with unique fashion recommendations based on their purchase history from 2018-2020, along with supporting metadata. We were motivated to work on this project because it would provide each customer with a personalized suggestion to enhance their shopping experience, reduce shopping time, and increase business sales. We attempted to suggest articles to customers using the findings of our analysis, rules for associations between articles, and data clustering based on age groups.

A. *Dataset Description and Pre-Processing*

The datasets used are from one of the kaggle competitions[1]. The dataset is divided into three sub-parts: articles metadata, customers metadata and transaction data.

The article's metadata consists of 25 features and 105k rows. This dataset is further divided into categories based on color, product type, gender etc.On examining the datatype of each column, 11 features contain numerical data and 14 features contain categorical data. On further scrutinizing the data based on the columns, we figured that the numerical columns could also be expressed as categorical types but with numeric labels and using which we can claim the data duplicacy in the columns. Data duplicacy is removed by dropping the columns and new preprocessed data has only 18 columns.

The customer's metadata consists of 7 features and 1.3M rows and contains information about their club membership status, postal codes, age, etc., which helps us analyze buyer's purchasing behavior. Based on the age range of the customers we tried to create groups using unsupervised algorithms.

The transaction dataset was one of the biggest dataset with 31M rows and 5 columns. First, in order to load the data into the Kaggle kernel we had to use the Kaggle's P500 GPU along with the cudf library which uses GPU clusters in order to load the chunks of dataset into memory. In addition we have to perform the following preprocessing steps in order to reduce the disk space occupied:

1. Converting article_id & customer_id into int32 type from char type, by defining appropriate types pandas reduces the space occupied by the columns.
2. Filtering out the year 2020 from the rest of the data as taking the most recent year for analysis would have a higher impact on the results.
3. Converting the dataset file from .csv to .parquet which takes smaller disk space and the read-time is lesser for parquet files.

II. Exploratory Data Analysis

Exploratory Data Analysis gives us key insights about the dataset. H&M has provided us with a lot of data and many interrelated associations can be formed between our customers and the articles that they purchase. Some of the key questions we hope to answer are what, why and where? What are our customers purchasing? Why are they purchasing those products, is there a trend? Inorder to get these insights we started with performing EDA on the 3 individual datasets and then merging them and getting combined insights from data as a whole. We will go through some things that stuck out from the data and which will help in some of the business decisions.

A. Articles Dataset

Article data set contains metadata about each article like their color value, pattern, people group, product type etc. We used this dataset to explore more about H&M products whether they try to target some specific gender or they produce more fancy clothings compared to subtle color clothes.

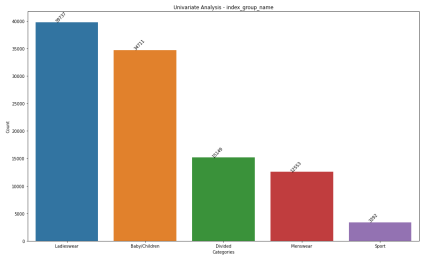


Figure 1.1: Count plot of Index Group Name

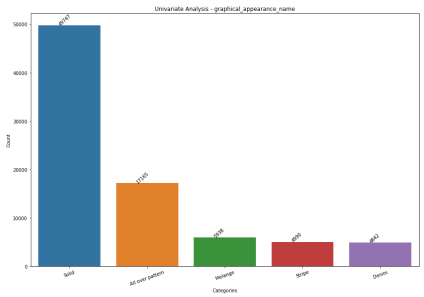


Figure 1.2: Count plot of Graphical Appearance

On examining the above two column data, most of the articles, approximately 38% belong to the Ladieswear group followed by babywear,divided, Menswear and sportswear. So we can conclude that around 70% of H&M articles are Ladieswear and Babywear and the rest of them are for other groups. Counting the articles based upon their graphical appearance (Solid, All Over Pattern, Denim etc.), 47 percent of articles have Solid graphical appearance. Figure 1.2 only represents the top 5 categories among 30 of them. Also around 59% of ladies apparel have Solid graphical patterns.

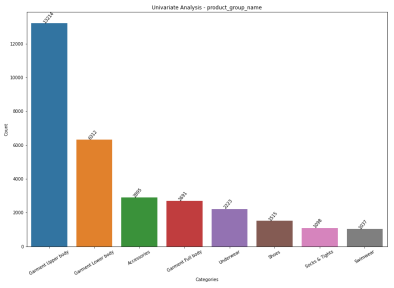


Figure 1.3: Count plot of Product Group Name

Next we tried to go over product group name(top 8 categories from total 19 categories) information and can deduce that major data recorded was upper body followed by lower body garments and Accessories. Among all ladieswear 26 percent of articles are Upper wear garments and only 14 percent ladieswear belong to Lower body garment product type.

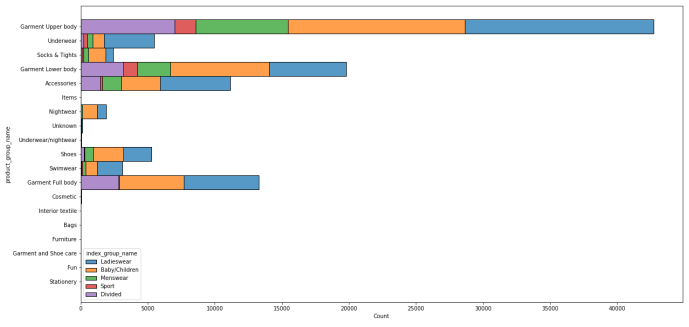


Figure 1.4: Product Group Name V/S Index Group Name

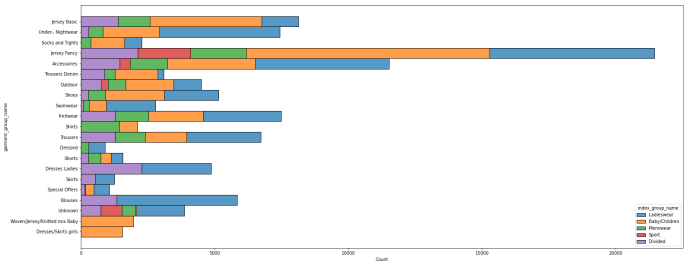


Figure 1.5: Garment Group Name V/S Index Group Name

From the bivariate analysis of garment group name and color value name column we found out that most garments had a darker tone irrespective of their group categories followed by light shade. From the second graph visual we were trying to find the relationship between garment group and index group columns, and can conclude that most of the garment groups have a maximum number of articles for either ladies or baby except for shirts, which have maximum articles for men.

Figure 1.6: Word Cloud of Description column

Word Cloud is used to visually represent the most frequent words written for the article's description. Pipeline to build word cloud starts with textual data pre-processing that involves stop word removal, punctuation removal followed by tokenization and lemmatization of the text. In Figure 1.6 bigger and bolder text represents most frequently used words.

B. Customers Dataset

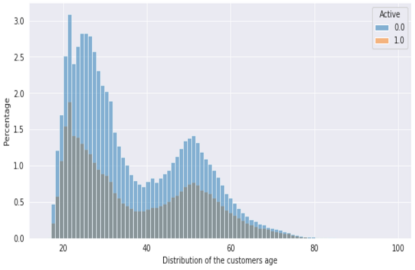


Figure 1.7 - Distribution of customer ag with percentage

H&M customers are widely spread through all the age groups right from age 20 to age 80. We see a large number of our buyers who are teenagers and mid-age range (two peaks in the plot). Knowing customer's ages can help us provide recommendations which are tailored for every age-group.

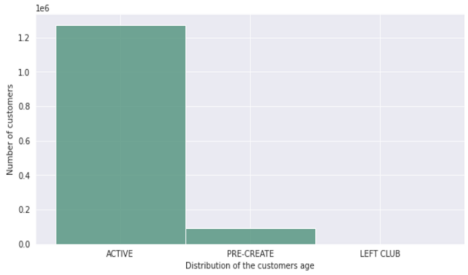


Figure 1.8

The above plot represents the count of customers in each club member status. Active club member status has the maximum number of members followed by pre-create and left club.

C. Transaction Dataset

The Transaction dataset was one of the key datasets, containing a million rows. With only 5 columns, transactions contain a lot of information about every purchase. Each row of transaction can be linked to a customer and his purchase.

	t_dat	customer_id	article_id	price
0	2018-09-20	-6846340800584936	663713001	0.050831
1	2018-09-20	-6846340800584936	541518023	0.030492
2	2018-09-20	-8334631767138808638	505221004	0.015237
3	2018-09-20	-8334631767138808638	685687003	0.016932
4	2018-09-20	-8334631767138808638	685687004	0.016932

The first question that comes to mind when looking at a transaction is what is the most bought item at H&M. Our analysis [fig2.0] shows that *Trousers* are the most bought item by our buyers followed by other popular candidates such as t-shirts, *shorts* and other apparels. It would be interesting to know why trousers are most popular among the others.

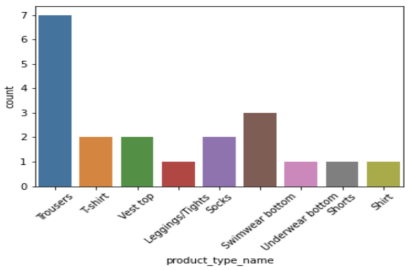


Figure 2.0 - Distribution of products and purchase count

So if *trousers* are most bought does that mean the lower garments are the most promising product category for H&M? It turns out yes!! lower garment category brings in the most sales to H&M, and we can visualize the distribution of these categories [fig 2.1]

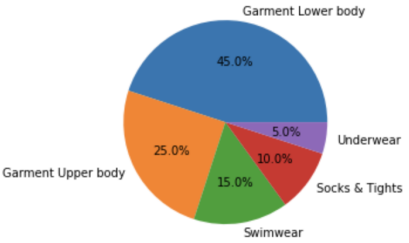


Figure 2.1 Distribution of product categories at H&M

Yet another important aspect of transactions is sales. What sales does each product bring in and if *jeans* are the most selling product does jeans sales are off the charts? Turns out not quite, as per [fig 2.2] even though *jeans* was the most selling product, *slacks* has the most sales.

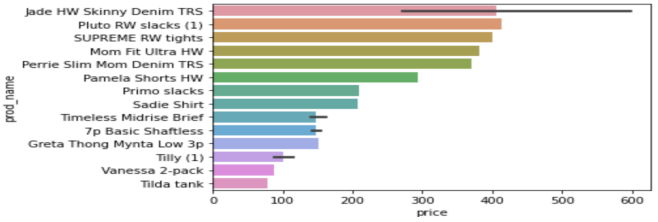


Figure 2.2 Top sales by product

III. Results and Discussion

A. *Recommendation using the apriori algorithm*

Generating itemsets: First step is to generate itemsets for every customer, we grouped all the item id for every customer to generate itemsets.

	customer_id	article_id
0	-9223343869995384291	[791142001, 821759001]
1	-9223279922255655589	[573716012, 835348001, 816832010]
2	-9223251502792607675	[679284001, 861024001, 552716001, 794819001]
3	-9223020246005719412	[856527001, 798050001]
4	-9223014153176581410	[201219013, 570004108, 570004112]
...	...	...
218756	9222665711703637549	[823165005, 823118004, 868018002, 875350001]
218757	9222733909628763172	[892910002, 898412001, 832473003]
218758	9222844312705255886	[832114005, 599580044]
218759	9222940818992675193	[918292011, 720125007, 720125001, 769748014, 8...
218760	9223148401910457466	[859101009]

Generating support values for each itemsets:

We apply the apriori algorithm implemented by the mlextend library to generate the support rules for our associations generated using the aprori most frequent itemsets principle.

	support	itemsets
27	0.003044	(806388002)
28	0.003067	(817354001)
4	0.003090	(557599022)
6	0.003127	(610776001)
36	0.003204	(895002002)
22	0.003209	(762846006)
37	0.003223	(896152002)
35	0.003236	(884319001)
23	0.003259	(768912001)
14	0.003310	(717490008)
25	0.003319	(791587001)
10	0.003332	(706016002)
5	0.003351	(579541001)
42	0.003355	(915529001)
26	0.003419	(806388001)

B. *Recommendation using our EDA:*

Our EDA analysis generated a lot of insights about our customers and their purchases. Our analysis showed some patterns of products that people buy together based on the color or to complete the attire. If one buys a light color shirt he would also buy a dark color jacket with darker pants. Similarly customers also buy matching socks or other utilities along with their clothing options. By visualizing the products[fig2.3] that customers purchase together, we create association rules of these products which are bought together and recommend the other products in the category. Lets see at some of these associations from our data:

Figure 2.3 Plot of most common product bought together

C. *Recommendation using clustering:*

Inorder to cluster and visualize the data we had to lower the dimensions of data to 2 dimensions. Here we used age as the grouping factor and clustered our customers based on their age groups and the type of product group that they purchase from. Figure 2.4 shows the 4 evident age groups were formed and the associated product groups they buy.

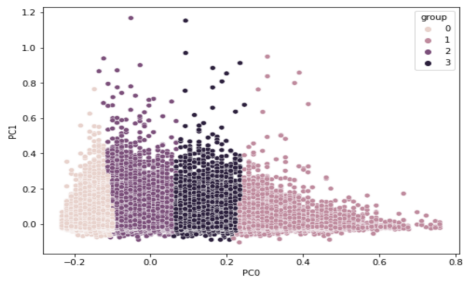


Figure 2.4 Results of K-Means clustering algorithm on age group and product groups.

D. *Final Recommendation:*

Our final recommendation is a merging of all our previous 3 recommendations for each of our customers and then creating a list of all articles for every customer that we think he is likely to purchase. We created a submission.csv file with 2 columns customer id and article id, each row representing every customer and associated list of article id separated by spaces.

	customer_id	prediction
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	0706016001 0706016002 0372860001 0610776002 07...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	0706016001 0706016002 0372860001 0610776002 07...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	0706016001 0706016002 0372860001 0610776002 07...

IV. Future Scope

We can implement classification algorithms to classify product categories given different product attributes.

In addition, by using historical data from 2019 and 2018, we can create more association rules and better recommendations. Additionally, neural networks can be utilized to :-

- Develop models to predict the price of a product based on several product attributes.
- Generate labels for articles and images using complex neural networks.
- Use the product description to generate product images using transformers and GANs.

V. References

1. Product Recommendation Systems using Apriori in the Selection of Shoe based on Android
2. Review of clustering-based recommendation systems.
3. *https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data*
4. *https://www.kaggle.com/code/datark1/detailed-eda-understanding-h-m-data/notebook*
5. *https://paperswithcode.com/paper/collaborative-similarity-embedding-for*
6. *Deep learning based recommendation system - SHUAI ZHANG, LINA YAO*
7. *https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/*
8. *Notebooks shared by Professor on course website*