Carnegie Mellon University

Personalized Product Prediction in Fast Fashion Industry

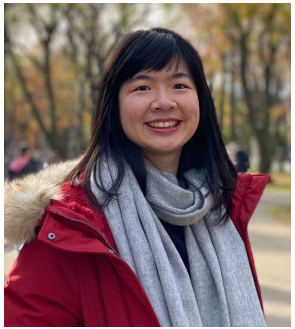
H&M Dataset

Presenters

Team 6

Eshan Mehrotra

MISM-BIDA 16



Angel Lee

MISM-BIDA 16

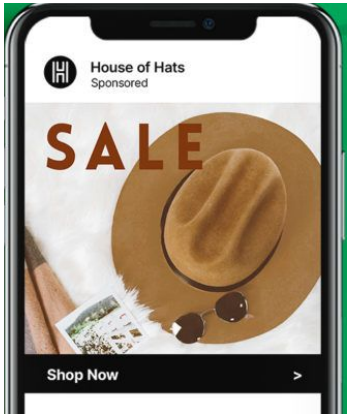
Carnegie Mellon University

Key Findings

- The model could help **identify specific groups of customers with high preference for specific product features.**
    - E.g. A specific group of customers (7.5%) visit H&M for Ladies swimwear and beach products that is preferably Black or White in color .
    - Another group (17.5%) visit H&M's website primarily for black or blue underwear / nightwear.
- Age was not a differentiating factor when it came to customer segmentation based on product preferences. Each segment had age ranges around 17-82 with a mean of ~36-37 years.

Motivation

- Nowadays, brands produce thousands of products and it is challenging for companies to pick the right product to market to each customers.
- We would like to investigate in the E-Commerce industry and understand how we could provide personalized experiences. Therefore, we would like to deep dive into people's purchasing behaviors and develop models that could help this domain.



Carnegie Mellon University

Key Datasets

- **H&M** released a Kaggle challenge and provided **transaction data for year 2018, 2019, 2020**.
- The datasets included **4 major files** and a folder:
  - articles.csv
  - customers.csv
  - transactions_train.csv
  - sample_submission.csv
- Between the 3 non-submission files, there are a total of 39 columns and 33 millions rows ((105542 for articles, 31788324 for transactions, and 1371980 for customers).
- Some of the interesting columns that we plan to explore include the ones with **details about the articles description**, **their color and perceived color, the product and department names, the customer age, their preferred buying method (online or offline), date of purchases, purchasing price** etc.

Carnegie Mellon University

Project Summary

Problem

- The primary pain point of fast fashion brands currently is their Year-on-Year **increasing advertising spend**. For example, H&M spent just under $100 million on advertising in the past year with close to $25 million on just digital ads.
- Even though online sales increased since covid, this **digital marketing spend was not able to be translated into sales** with a 21% drop in YoY sales in H&M's first quarter of 2021.

Project Goal

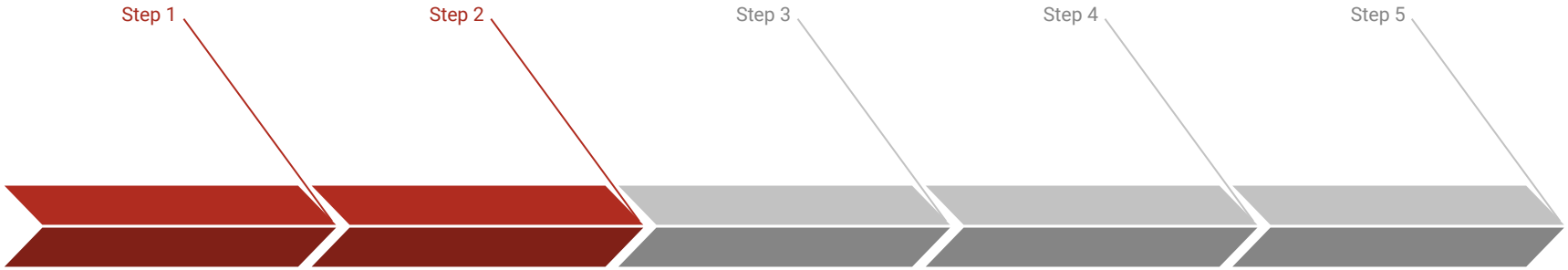
Cluster customers to predict top selling products for each cluster based on their preferences.

Value-add to Stakeholders:

- Marketing team could **allocate their budget more efficiently** and **have better return-on-ads-spend** by advertising the top predicted product for each segment to the customers in that group.
- On the launch of a new product, based on the product features and segment preferences, H&M can do **targeted marketing** to the specific segment thereby **decreasing their huge marketing spend**.

Carnegie Mellon University

Project Methods



Sampling Dataset

Taking a smaller sample to focus on 13,633 customers

Data Cleaning

Remove outliers and undesirable feature values

Feature Engineering

Selecting desired product features, vectorizing on basis of frequency per customer to prepare for model ingestion

Segmenting Customers

Using transaction history to understand purchasing behaviour based on product characteristics

Product Prediction

Interpret each customer cluster based on product preferences and list out top product recommendations matching those preferences.

Carnegie Mellon University

Sample Dataset to focus on subset of Customers

```
Total Number of Customers in Entire Training Data: 1371980
Total Number of Customers in ~1% of Data: 13633
```

Fig 1 Unique Customers

Data Cleaning

	customer_id	age	product_group_name	perceived_colour_master_name	index_group_name	garment_group_name	sales_channel_name
315945	677e227cf04d0b729ba4d28fa33c96af1e64105beb3cb5...	31.0	Socks & Tights	undefined	Ladieswear	Socks and Tights	Online
316318	fcfd6da6a507ef25b079b61a355b0ef57a09a63fae962c...	29.0	Garment Upper body	Unknown	Menswear	Jersey Basic	Store
316370	fcfd6da6a507ef25b079b61a355b0ef57a09a63fae962c...	29.0	Garment Upper body	Unknown	Ladieswear	Jersey Fancy	Online
316686	5b7c093e10aad6ef556044628984a214e2bbaafff42565...	51.0	Garment Upper body	Unknown	Ladieswear	Jersey Fancy	Store

Fig 2 Feature Set of interest

Column	Undesired Values Count
product_group_name	893
perceived_colour_master_name	2937
index_group_name	0
garment_group_name	3711
sales_channel_name	0

Fig 3 "Unknown" or "undefined" value counts

```
Total Number of Customers after cleaning data 13573
```

Fig 4 Unique Customers post cleaning

Carnegie Mellon University

Feature Engineering

Raw Format after Feature Selection

	customer_id	age	product_group_name	perceived_colour_master_name	index_group_name	garment_group_name	sales_channel_name
33977	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Lower body	Black	Divided	Trousers	Online
33978	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	Black	Ladieswear	Jersey Basic	Online
33979	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	Pink	Divided	Knitwear	Online
33980	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	Grey	Ladieswear	Knitwear	Online
33981	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	Grey	Ladieswear	Knitwear	Online
33983	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	Beige	Divided	Knitwear	Online
33984	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Underwear	Black	Ladieswear	Under-, Nightwear	Online
33985	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Lower body	Blue	Divided	Trousers	Online
33986	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Lower body	Grey	Ladieswear	Jersey Basic	Online
33988	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	Garment Upper body	White	Ladieswear	Jersey Basic	Online

Fig 5

Vectorized form to feed into model

	customer_id	age	Online	Store	Baby/Children	Divided	Ladieswear	Menswear	Sport	Beige	Black	Blue	Bluish Green	Brown	Green	Grey	Khaki green	Lilac Purple
1420	39725c51fec5cf7edeb4e3885cf0dc42db731ac7bc5151...	32.0	10	0	0	4	6	0	0	1	3	1	0	0	0	3	0	0

Fig 6

Carnegie Mellon University

Cluster Customers and assign groups respectively

```
Customer:  0797fbc3fb242be600e8a707ef9976201bc325c02cb1048178cc7d1ff299ffa7
Topic  0 :  0.3351604720534883 %
Topic  1 :  34.6257586222514 %
Topic  2 :  0.33482147911178184 %
Topic  3 :  64.36963554185057 %
Topic  4 :  0.3346238847327544 %
```

Fig 7: Probabilities of belonging in each segment for a customer

```
Customer:  bf2ce07a1756eb835d6ecfca1c53a6a8038b743866efbd9462578d1a778e4042
Topic  0 :  12.686990961779987 %
Topic  1 :  22.16969725510996 %
Topic  2 :  0.1786148659428504 %
Topic  3 :  17.683534004225816 %
Topic  4 :  47.28116291294139 %
```

Fig 8 Probabilities of belonging in each segment for different customer

Segment	Number of Customers
0	2372
1	1375
2	1029
3	4926
4	3871

Fig 9: Segment assignments based on highest probabilities for all customers

Carnegie Mellon University

Reference: Analyzing the Browsing Basket: A Latent Interests-Based Segmentation Tool

Product Prediction: for each customer cluster, list out top product with selected characteristics to recommend

```
[Topic 0]
Under-, Nightwear : 0.1323581251840241
Online : 0.1235887621528515
Underwear : 0.11953748283381468
Ladieswear : 0.08812733753896582
Baby/Children : 0.08750515766507515
Black : 0.0529032531634284
Jersey Fancy : 0.04155896216780223
Garment Upper body : 0.03906568130510786
Blue : 0.03549805799174251
Grey : 0.02777001980053558
```

	article_id	prod_name	count
0	0885951001	Calypso C&S push bralette	766
1	0923037001	Hongkong	638
2	0758034001	Ivory ch brazilian 5pk	615
3	0923340001	Lion top	598
4	0903773001	Minami	574
5	0865926002	Sonja top	567
6	0909912001	LW (K) Carla cons Hoodie	557
7	0926166001	LW (K) CATRIN consc cardi	528
8	0719655001	Greta Ch hipster ctn 3p	515
9	0611415001	Charlotte Brazilian Aza.Low 2p	506
10	0865929007	Minja 2.0	463
11	0918894001	Anakin body	421

```
[Topic 2]
Online : 0.18345189888239274
Ladieswear : 0.18116629585753669
Swimwear_x : 0.12275345434229956
Swimwear_y : 0.12081283760687203
Black : 0.0725871058795002
Garment Upper body : 0.03077088965122557
White : 0.029660973831592982
Blue : 0.019807852244858763
Garment Lower body : 0.01835805538866522
Divided : 0.01675840690839869
```

	article_id	prod_name	count
0	0812668001	B Lola Beach Dress	131
1	0866383006	Push it Push Bra	130
2	0806225002	All That Jazz Push Up Bra	117
3	0811899003	X-tina swimsuit	115
4	0822946001	Scary Spice Top	110
5	0739590027	Timeless Cheeky Brief	101
6	0559633008	Jane Swimsuit	95
7	0822959001	Scary Spice Swimsuit	92
8	0873072002	X-tina long top	87
9	0812364013	X-tina Brazilian	87
10	0882759003	Price Mini Curved Brief	84
11	0599580072	Timeless Midrise Brief	83

Carnegie Mellon University

Figure 10: Top 10 characteristics of Products in Segment 0 and 2 with the predicted Top 12 products

Future scope

- **Time Series Analysis:** consider seasonality of the fast fashion industry
- **Scale up dataset size**: Scale up the data used to larger number of customers to check robustness of the method and approaches
- **Test different approaches of communicating**, there are different channels, e.g. email, app, in-store, and different timing, e.g. weekend vs weekdays, morning vs evening, that we communicate with our customers. We could track which combination could better boost transactions

Thank you!