**Leveraging Data Science for Growth: H&M**

**Sydney Ng Deng Jun, Bhairavi Vairavelu, Dai Yu, Widya Tantiya Yutika, Yin Hang**

**Associate Professor Zhu Feida**

School of Computing and Information Systems

Singapore Management University

**1. Overview**

**1.1. Introduction**

H&M is a Swedish multinational fashion retailer that has been in business for several decades. This company has grown to become one of the largest fashion retailers in the world and is best known for its trendy and affordable clothing. While having a strong presence in the industry, H&M is one of the world’s top 10 most valuable apparel brands in 2022 [1]. As of November 2022, H&M operates in 78 geographical markets with a total of 4465 stores worldwide [2]. This company has employed over 100,000 people, making it one of the largest employers in the retail industry [3].

However, H&M has faced several challenges in recent years. One of the main challenges is increasing its sales and driving growth in an extremely competitive market. This company is looking to improves its sales figures and meet its targets for 2023 despite facing challenges related to supply chain disruptions and inflationary pressures [5]. H&M has also been working to counteract the forecasted impact on products to increase customer growth.

In this report, we will examine the challenges faced by H&M and how we can leverage data science concepts to address these challenges.

**1.2 Objectives**

To succeed in the growingly competitive retail industry, H&M must target customers more effectively to improve revenue and boost transactions. As such, our team has identified several objectives for this project to address the challenges that H&M is currently facing, as well as the data science concepts that will be used to increase sales and drive growth for H&M.

The first objective is to conduct an RFM analysis which will provide insights into customer behaviour pattens and preferences. By analyzing customer’s recency of visits, frequency of purchase and the monetary value of their purchase, H&M can segment its customer base and potentially create targeted marketing campaigns that will cater to each segment’s specific needs.

Another objective of this project is to conduct a market basket analysis to identify purchase patterns among customers. With this insight, H&M will be able to identify cross-selling opportunities and can create bundled promotions and product recommendations that can boost their sales and transactions.

The third objective is to develop a personalized product recommendation system that will enable H&M to promote specific products to individual customers based on their purchase preference and behaviour. This will in turn help H&M to create a personalized shopping experience for each customer, which can increase customer satisfaction and loyalty.

Overall, by achieving these objectives, H&M will be able to improve their customer engagement and loyalty, and ultimately drive sales growth for H&M. By gaining a deeper understanding of customer behaviour, as well as their patterns and preferences,

**2. Data Collection and Preparation**

**2.1 Data Source**

We are using H&M purchase history of customers across time data from kaggle.com:

*https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data*

From the given dataset, we have selected the below 3 datasets for further analysis:

1. **Articles**

This dataset includes items available for purchase in H&M and their detailed information. The below columns/ parameters are selected for analysis purpose.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Assumption** |
| article\_id | Unique identifier for each H&M article | - |
| prod\_name | Name of the product | - |
| prod\_group\_name | Product group section name | Categorized into 19 product groups |
| graphical\_appearance\_name | Pattern of the products | - |
| colour\_group\_name | Color of the products | Represent major color |
| index\_group\_name | Index group | Categorized into 5 index groups i.e. Ladieswear, Baby/Children , Menswear, Sport and Divided |

1. **Customer**

This dataset includes customer detailed information. The below columns/parameters are selected for analysis purpose.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Assumption** |
| customer\_id | Unique identifier of each H&M customer | - |
| club\_member\_status | Customer status of membership | Null value refers to non-member |
| fashion\_news\_frequency | How often the customers want to get news from H&M’s Fashion Newsletter | Null value refers to not subscribing to H&M newsletter |
| age | Age of customers | Since there is a lot of missing values, age will not be further analyzed |

1. **Transactions Train**

This dataset consists of the purchases done by each customer over time. The below columns/ parameters are selected for analysis purpose.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Assumption** |
| t\_dat | Date of transaction | Reformatting into datetime type |
| customer\_id | Customer ID for the transaction | - |
| article\_id | Article ID of products bought by customers | - |
| price | Price of products bought by customer | Price in 1,000 or k |
| sales\_channel\_id | The channels in which customer bought the products from | Sales Channel ID 1 refers to online purchases and Sales Channel ID 2 refers to in-store purchase |

**2.2 Data Preparation**

The first step on data preparation was to clean the data. This includes finding missing value and replace it, accordingly, removing duplicated rows, reformatting of the data. For articles dataset, there is no missing values on all fields except detail description. Since detail description is not used for our analysis, we will further remove the fields. Next step is to remove on duplicated row if exists. For article and customer dataset, there is no duplicated row found. However, there are several double entry data in transaction dataset which was removed from the datasets. Once all the datasets have been cleaned, it will be further processed and analysed.

**3. Analysis and Discussion**

In this section, we will elaborate how we conduct our analysis.

**3.1 *RFM Analysis***

The RFM analysis and customer segmentation are essential tools for H&M to understand its customers better and tailor marketing strategies to maximize profitability. This section will describe the process of conducting RFM analysis and customer segmentation and explain their importance for H&M.

***3.1.1 RFM Calculation***

To begin, we are using combined dataset of customer and transaction\_trains and import the data into a pandas Data Frame.

The RFM values refers to Recency, Frequency, and Monetary respectively. They are calculated for each customer. We first determine the snapshot date, which is the day after the latest transaction date in the dataset. The dataset is then grouped by customer\_id, and we calculate Recency –the difference between the snapshot date and the most recent transaction date for each customer, Frequency –the number of transactions for each customer, and Monetary –the total amount spent by each customer.

Customer who has not purchased any product with September 2018 to September 2020 will not be included in the analysis.

***3.1.2 Quantile Segmentation***

After calculating the RFM values, we determine quantiles for each metric. These quantiles allow us to assign R, F, and M scores to each customer based on their respective RFM values. Customers with lower Recency values receive higher R scores, while those with higher Frequency and Monetary values receive higher F and M scores, respectively.

We then combine the R, F, and M scores to calculate the total RFM score for each customer. Based on these scores, we define customer segments as below:

* Low Spender: total RFM score <= 6
* Promising: total RFM score <= 9
* High Spender: total RFM score <= 11
* Champion: total RFM score > 11

Customers are assigned to a segment according to their total RFM score.

***3.1.3 RFM Customer Segmentation Visualization***

To visualize the results, we create plots to display the distribution of Recency, Frequency, and Monetary values, as well as the count of customers in each customer segment. This visual representation helps H&M understand the distribution of its customer base and identify trends.

Graphical user interface

Description automatically generated

Finally, we display the results, such as the top customers based on Monetary value and the customer count in each segment, which can be used to guide our marketing strategies.

**Chart, bar chart

Description automatically generated**

**3.2 *Market Basket Analysis***

Market Basket Analysis is a statistical technique used by retailers to gain insights into customers' buying behaviour and make informed decisions about product placement and marketing strategies. It involves identifying patterns of co-occurrence of items in transactions.

The Apriori algorithm is a popular algorithm used to perform Market Basket Analysis. It works by identifying all the frequent itemsets in the dataset and then generating association rules that capture the relationship between the items in the frequent itemsets. Frequent itemsets are sets of things that occur together in many transactions.

In our implementation, we first converted the transactional data into a one-hot encoded format where each row represented a customer, and each column described an item. Next, we identified the frequent itemsets by selecting only those with support more significant than a specified minimum support threshold of 0.03. We then generated association rules that capture the relationship between the items in the frequent itemsets, using the lift metric to measure the strength of the association.

Finally, we filtered the association rules based on a minimum threshold for support and lift to identify the most interesting and actionable rules. We used the mlxtend library in Python to implement the Apriori algorithm and visualize the results using a heatmap. Additionally, we identified the top-selling products for each customer group.

Sorting by lift, support and confidence, pivoting by support, our heat map shows a confidence of 0.71 for the low spender, which means that given a transaction containing the antecedent item(s), there is a 71% chance that the same customer will also buy the consequent item(s). In other words, the rule "if A, then B" has a confidence of 0.71, indicating a strong association between A and B. Similarly, for the Promising segment, a confidence of 0.53 meant that given a transaction containing the antecedent items, there is a 53% chance that the same customer will also buy the consequent items. Conversely for the High Spenders and Champions, there was a confidence of 0.34 and 0.28 instead, suggesting that High Spenders or Champions usually had more diverse shopping habits.

|  |  |
| --- | --- |
|  |  |

**3.3 *Recommendation System***

Building on the foundation of the RFM and Market Basket Analysis, we sought to create a recommender system that would cater to customers who had previous transactions with us and could be segmented. To that extent, we made two functions, one for the first item recommendation and another for the following item recommendation.

The first\_item() function recommends the first item a customer should consider based on the product groups they have previously purchased. It does this by looking at the antecedents of the association rules generated through market basket analysis. The function then retrieves the product group names from the association rules, removes duplicates and recommends the top articles associated with those product groups.

The second\_item() function recommends the following item a customer should consider based on the product groups they have already added to their cart. It checks the antecedents of the association rules generated through market basket analysis and compares them to the product groups in the cart. If there is a match, it retrieves the consequents of that association rule, removes duplicates and recommends the top articles associated with those product groups. If there is nothing in the cart, it calls the first\_item() function to recommend the first item. If there is only one item in the cart, it retrieves the top articles associated with the product group of that item.

Lastly, we created a recommender function to improve customer experience; The process recommends items to customers based on past transactions. The method first prompts the user to input their customer ID, which is taken to be checked against the customer list. If the ID matches a customer, the program identifies which customer group the user belongs to based on their past purchases and displays the group. However, if the user has never made a purchase, the program informs them that no recommendation is available.

On the other hand, if the user has a previous transaction, we will use market basket analysis to suggest items that the user can add to their cart. The program then generates a list of all possible combinations of these items and then tells a second item based on the customer group's past purchasing behaviour. This way, we can maximize the items already in the user's cart while offering additional items based on their interests.

**4. Limitations**

Bias needs to be a concern in the project as it can have a significant impact on the accuracy and fairness of the recommendations provided to the users before the implementation. bias can arise in various forms, such as content bias, algorithmic bias, and user bias. Without address the Bias, the users may be consistently recommended items that are not relevant to them, while others may be excluded from relevant recommendations. This can lead to a loss of trust in the recommendation system, and ultimately result in decreased engagement and revenue for H&M.

**5. Data Governance**

Data governance is important for this project as it helps ensure that the data used in the system is accurate, reliable, and transparent, and that it is processed in a way that protects customer privacy.

There are several potential data governance issues that we need to look into:

1. Privacy: The transaction data may contain personal information. If this data is not properly secured, it could lead to breaches of privacy and other legal consequences.
2. Data quality: The transaction data needs to be of high quality to ensure accurate results.
3. Data accessibility: The access to the transaction data needs to be restricted to only authorized personnel who have a legitimate business need for it. This can help prevent unauthorized access, misuse.
4. Compliance: There are several data governance laws to be followed, such as Personal Data Protection Act (PDPA) and Personal Data Protection Commission (PDPC) Guidelines. Non-compliance with these laws and regulations could result in legal and financial consequences for the company.

It is important for H&M to ensure that proper data governance measures are in place to address these issues and ensure the responsible and ethical use of customer transaction data.

Data Provenance is an essential part of data governance by documenting its data management processes and policies. It helps the organization in following aspects:

1. Improve Accountability: Documentation ensures that all decisions made during the development and implementation of the recommendation system are traceable, allowing for accountability in case of errors or malfunctions.
2. Ensure Trustworthiness: Documenting data sources, processing, feature selection, and model training provides transparency about the inputs and processes used to generate recommendations, which helps customers trust the system.
3. Increase Customer Loyalty: By providing transparent and trustworthy recommendations, H&M can build customer loyalty.

In this project, the data provenance is carried out in the following steps:

1. Data sources should document the data used to train and test the recommendation system, including customer behavior, such as purchase history, browsing behavior, and search queries, and product data, such as descriptions, images, and attributes.
2. Data processing should document the steps taken to process and clean the data, including any transformations or aggregations applied to the data.
3. Feature selection should record the features used in the recommendation system, including how and why they were chosen.
4. Model training should document the algorithms used for training the recommendation system, the hyperparameters chosen, and the evaluation metrics used to assess the model's performance.
5. Model evaluation should document how to evaluate the performance of the recommendation system, how to choose the metrics, and the limitations or assumptions that were made.
6. Ethical considerations should document any ethical considerations that were taken into account when developing and implementing the recommendation system, such as data privacy, fairness, and bias.

**6. Closed Loop Analysis**

To ensure the effectiveness of the recommendation system, H&M needs to collect feedback from customers and analyze new data. By monitoring the system's performance, H&M can identify areas for improvement and make changes to the system over time. For example, the company can analyze which recommendations customers are clicking on and purchasing and use this information to adjust the algorithm. We can also look into conversion rate or additional revenue generated to measure the performance and make adjustments to improve it over time.

**7. Conclusion**

In conclusion, our project aims to conduct an analysis using RFM analysis to understand H&M's customers' buying patterns. From this, we plan to build a tailored product recommendation system using association rule mining methods based on specific features to make personalized and relevant product recommendations to customers. Additionally, by implementing these data science methods, H&M can focus on products which increase sales, and customer loyalty, improve transaction quantity and increase revenue for H&M. To evaluate the effectiveness of the recommendation system, we can collect feedback and new data to analyze its performance and adjust to improve the system over time.

A personalized product recommendation system using advanced data science methods, which can help H&M increase sales, customer loyalty, transaction quantity, and revenue. By analyzing customers' buying patterns using RFM analysis, H&M can gain insights into what products customers are interested in and provide them with personalized recommendations.

Continuous improvement of the recommendation system is essential to keep up with changing customer preferences and trends. By consistently evaluating and improving the system, H&M can maintain its competitive position in the market, enhance customer satisfaction, and drive business growth.

**8. References**

[1] H&M Group, (2023) <https://hmgroup.com/about-us/markets-and-expansion/market-overview/>

[2] P. Smith, (Feb 16, 2023) Number of employees at H&M worldwide from 2005-2022

<https://www.statista.com/statistics/267375/number-of-hundm-employees-worldwide/>

[3] P. Smith, (Aug 15, 2022) Most valuable clothing companies in the world, by brand value 2021-2022

<https://www.statista.com/statistics/1207840/top-apparel-clothes-brands-worldwide/>

[4] ARHI KIVILAHTI, (JAN 31) H&M struggling to match the growth of it’s fast fashion peers

<https://www.adainsights.fi/blog/hm-struggling-to-match-the-growth-of-fast-fashion-peers>

[5] R. Anna (Sep 30, 2021) <https://www.reuters.com/business/retail-consumer/hms-june-august-profit-rises-more-than-expected-2021-09-30/>