Optimization of Product Merchandising and Sales Enhancement through Recommendation Systems Using Collaborative Filtering

Name: Md Tauhidul Islam
Id: a1895813

The University of Adelaide

4433_COMP_SCI_7306 Mining Big Data
Lecturer: Dr. Alfred Krzywicki

Table of Contents

1. Executive Summary

Project Overview

The project named "Optimization of Product Merchandising and Sales Enhancement through Recommendation Systems Using Collaborative Filtering," aims to transform the shopping experience both in-store and store website. This project uses sophisticated data analysis techniques to recommend products that customers are likely to purchase based on behaviours of the purchased items.

Problem Description

The grocery store management wants to improve sales and customer satisfaction by implementing a recommendation system so that they can suggest relevant products in-store and on their website. Product merchandising based on the recommendations could personalise the shopping experience as well as encourage users for larger purchases.

Benefits for the company:

Implementing this project will provide several key benefits:

- Personalised shopping experience: Personalised recommendations allow customers to find products quickly everything they need and the store to respect customer preferences, which eventually enhances customer satisfaction and loyalty.
- Increased sales: Recommending items based on the customers preferences may motivate them to add more items in their basket which may result in an increase in sales.
- Enhanced operational efficiency: Efficient merchandising and targeted recommendations may reduce the time and effort required to manage inventory and fulfill orders.
- Efficient marketing: The system may improve marketing strategies by focusing on products that are more likely to be bought by certain customers.

Feasibility and Scalability

The recommendation system is trained to efficiently to handle large volumes of data which makes it capable to accommodate approximately one million customer transactions. Therefore, the system will remain sustainable with the store’s growing customer base.

Test Results

The results from the recommendations systems can suggest top items for an individual customer as well as top items for a particular purchased item. The test results suggest that the system can efficiently suggest products to customers based on their preferences.

Conclusion

The integration of advanced collaborative filtering techniques into our recommendation systems could change the way the store engages with the customers. It has the potential to improve the store’s customer experience and sales significantly.

2. Introduction

In the retail industry, understanding customers' preferences and behaviour is significant for success due to its competitive nature and growing e-commerce demand. Customers might get tired of finding their preferred products from thousands of choices which is why personalisation of shopping experience is significant.

The goal of this project is to develop a system by which we can recommend products to customers based on similarities that can improve customer satisfaction by personalisation, improve operational efficiency and increase sales.

The scope includes developing a recommendation system using collaborating filtering that can identify the similarities in purchasing patterns. The model could propose new items based on the similarities to previously purchased items (content-based) or user's similarities to other users and their selected items (User-based). In order to group and compare items easily, content-based collaborative filtering has been implemented in this project.

Implementation of this recommendation system can significantly improve the merchandising techniques, improve customer engagement and enhance sales. The targeted audience for this recommended system is retail managers, marketing teams, or online shoppers.

3. Exploratory Analysis

The exploratory data analysis helps to understand basic properties of the data. The following analysis has been conducted as a part of this project. The analysis is significant for preparing the data for collaborative filtering techniques.

3.1. Loading the Data

The analysis starts with loading the transaction dataset into pandas dataframe. The dataset contains detailed records of customer purchases. The first step in the EDA was to visually inspect the initial few rows of the dataset to understand the available data fields.

	BillNo	Itemname	Quantity	Date	Price	CustomerID	cost
0	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850	20.34
1	536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	1/12/2010 8:26	4.25	17850	25.50
2	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850	15.30
3	536365	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850	20.34
4	536365	SET 7 BABUSHKA NESTING BOXES	2	1/12/2010 8:26	7.65	17850	15.30

Figure 1: 1st five items in the dataset

A brief description of all the columns of the dataset is as follows:

- **BillNo**: A unique identifier for each transaction
- **Itemname**: The name of the item purchased
- **Quantity**: The number of items purchased
- **Date**: The date and time of the transaction

- **Price**: Unit price for the item purchased
- **CustomerID**: A unique identifier for each customer
- **Cost**: The total cost for the item purchased

3.2. Data Structure

Further analysis involves reviewing the data types and non-null counts in our dataset. The below image reveals that we have 40,000 entries with no missing values in any of the 7 columns. The columns include a mix type of data where the quantitative discrete variables are integers, quantitative continuous variables are float type and object type refers to categorical variables.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   BillNo      40000 non-null  int64
 1   Itemname    40000 non-null  object
 2   Quantity    40000 non-null  int64
 3   Date        40000 non-null  object
 4   Price       40000 non-null  float64
 5   CustomerID  40000 non-null  int64
 6   cost        40000 non-null  float64
dtypes: float64(2), int64(3), object(2)
memory usage: 2.1+ MB
```

Figure 2: Some significant information of the dataset

However, the “Date” column is being considered as a categorical variable and this might not be useful for data visualisation. We will convert the “date” column into datetime format for visualisation purposes.

3.3. Statistical Analysis

Some key insights from the data can be derived from the numeric columns which includes mean, median, ranges, standard deviation and minimum and maximum values for each column.

	BillNo	Quantity	Price	CustomerID	cost
count	40000.000000	40000.000000	40000.000000	40000.000000	40000.000000
mean	540254.879225	3.487700	3.732165	15577.606525	11.097411
std	2380.444952	2.611766	5.711630	1730.347123	13.298301
min	536365.000000	1.000000	0.100000	12347.000000	0.140000
25%	538093.000000	1.000000	1.650000	14224.000000	3.300000
50%	540373.000000	2.000000	2.950000	15570.000000	7.950000
75%	542360.000000	6.000000	4.650000	17220.000000	15.800000
max	544398.000000	10.000000	295.000000	18283.000000	527.700000

Figure 3: Some significant statistics of the numeric columns

These statistics are crucial to understand distribution of quantities, prices, and costs, which informs the potential for product recommendations.

3.4. Data Preprocessing

As we saw that the "Date" column is not in the appropriate data type, we converted it to datetime format for visualisation purposes.

Moreover, we have removed the duplicate data from our dataset. We also removed users and items that have interaction below 5 and 10 times respectively for more efficient recommendation system as less interactive users or items have negligible impact on the model. The shape of the initial data was (40000, 7) where after filtering out duplicate and sparse data we have (33502, 7). We can see the differences in the below figure.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 33502 entries, 0 to 39999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   BillNo      33502 non-null  int64
 1   Itemname    33502 non-null  object
 2   Quantity    33502 non-null  int64
 3   Date        33502 non-null  datetime64[ns]
 4   Price       33502 non-null  float64
 5   CustomerID  33502 non-null  int64
 6   cost        33502 non-null  float64
dtypes: datetime64[ns](1), float64(2), int64(3), object(1)
memory usage: 2.0+ MB
```

Figure 4: Some significant information of the dataset after filtering

3.5. Data Visualisation

Frist, the figure 5 and 6 shows the number of transactions over time and day of the week respectively which can be significant for the store as it can be utilised to determine which time the year and day had the highest sales so that they can be prepared with recommendation for greater sales.

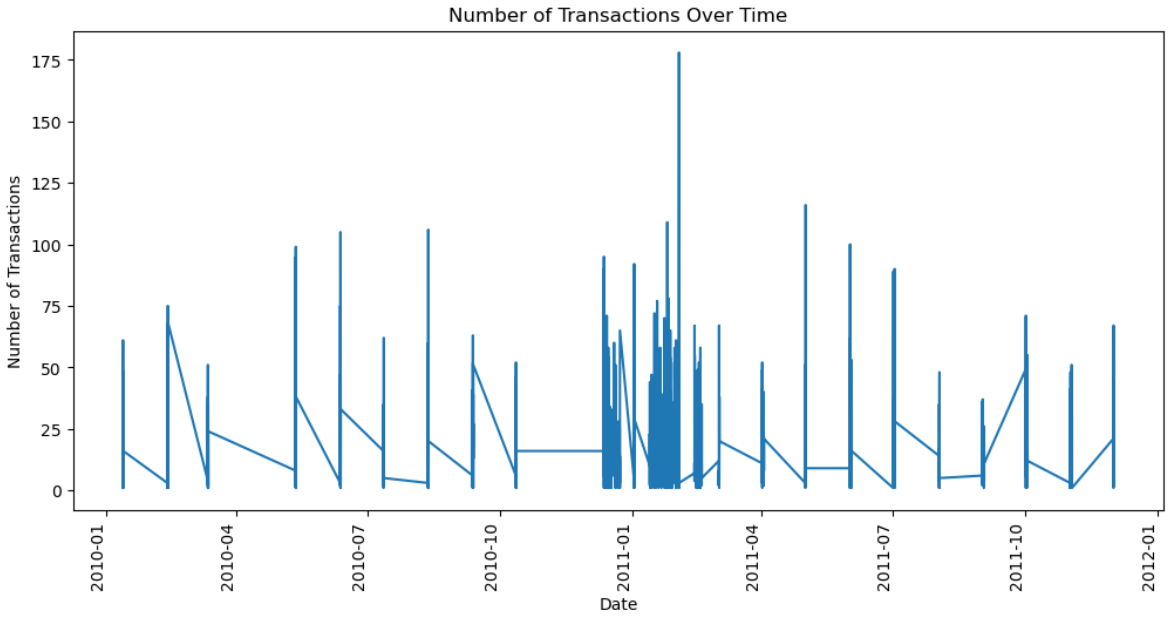

Figure 5: The number of transactions over time

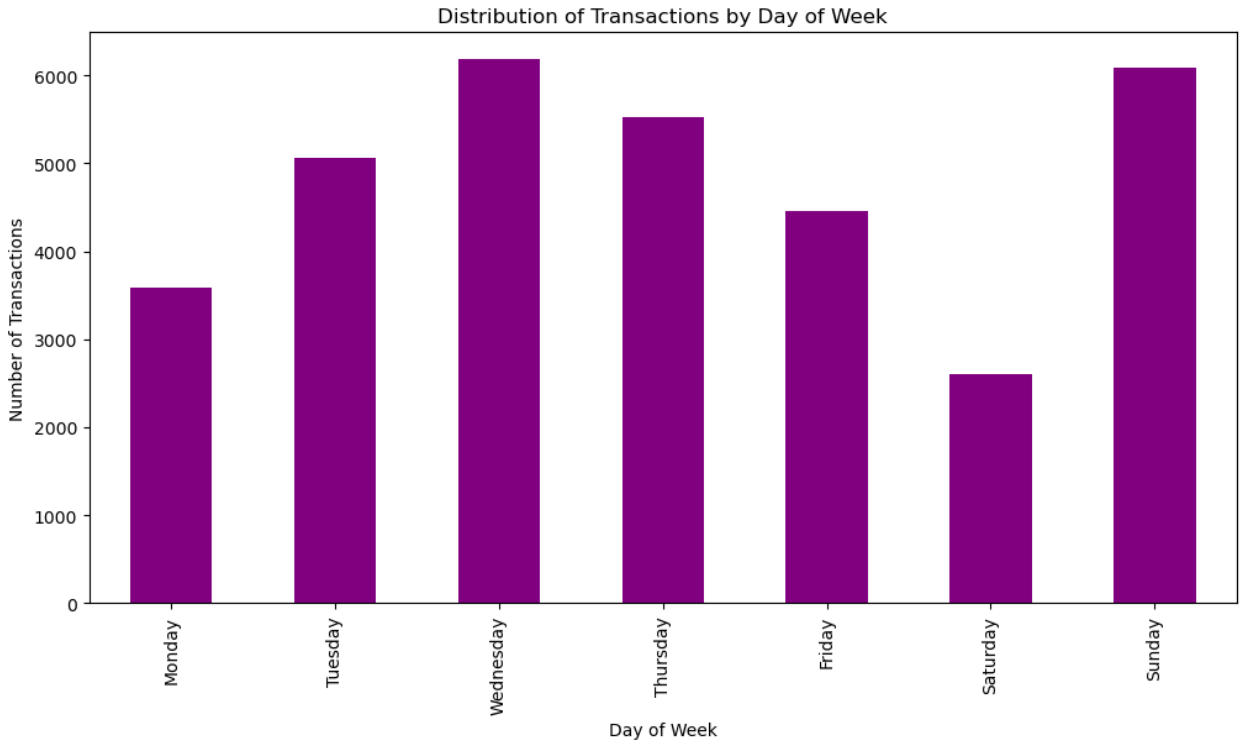

Figure 6: The number of transactions by day of the week

We also identified top 10 customers who are doing most purchases (Figure 7) and items (Figure 8) what are being purchased the most. This information will be used as an example in our recommendation system.

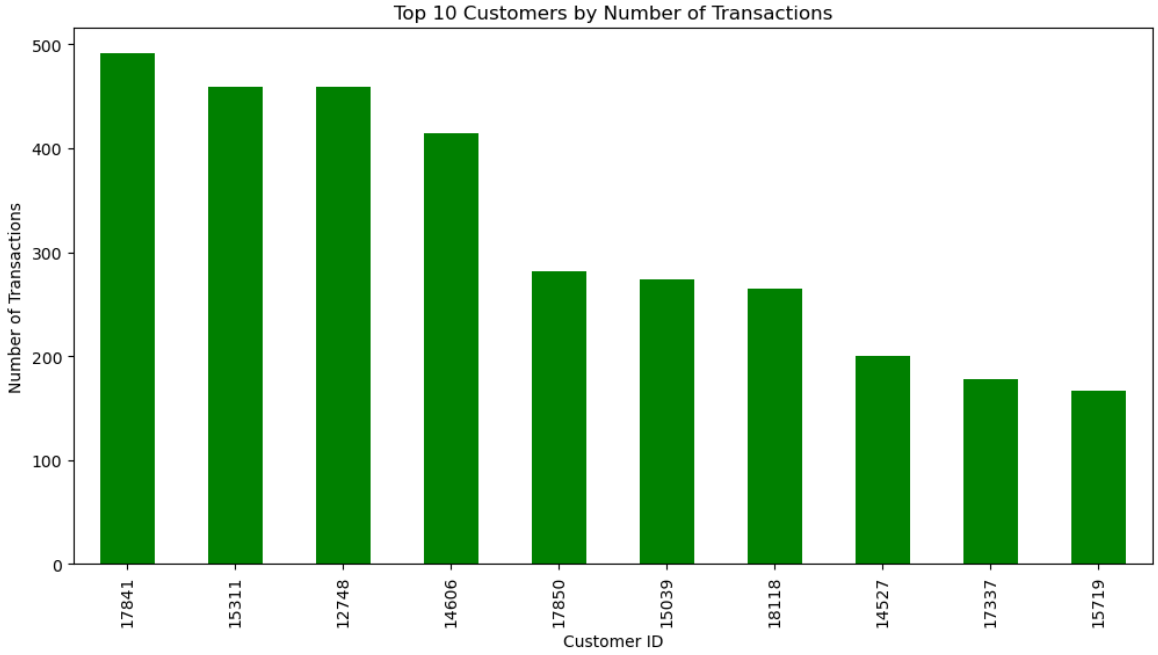

Figure 7: Top 10 customers

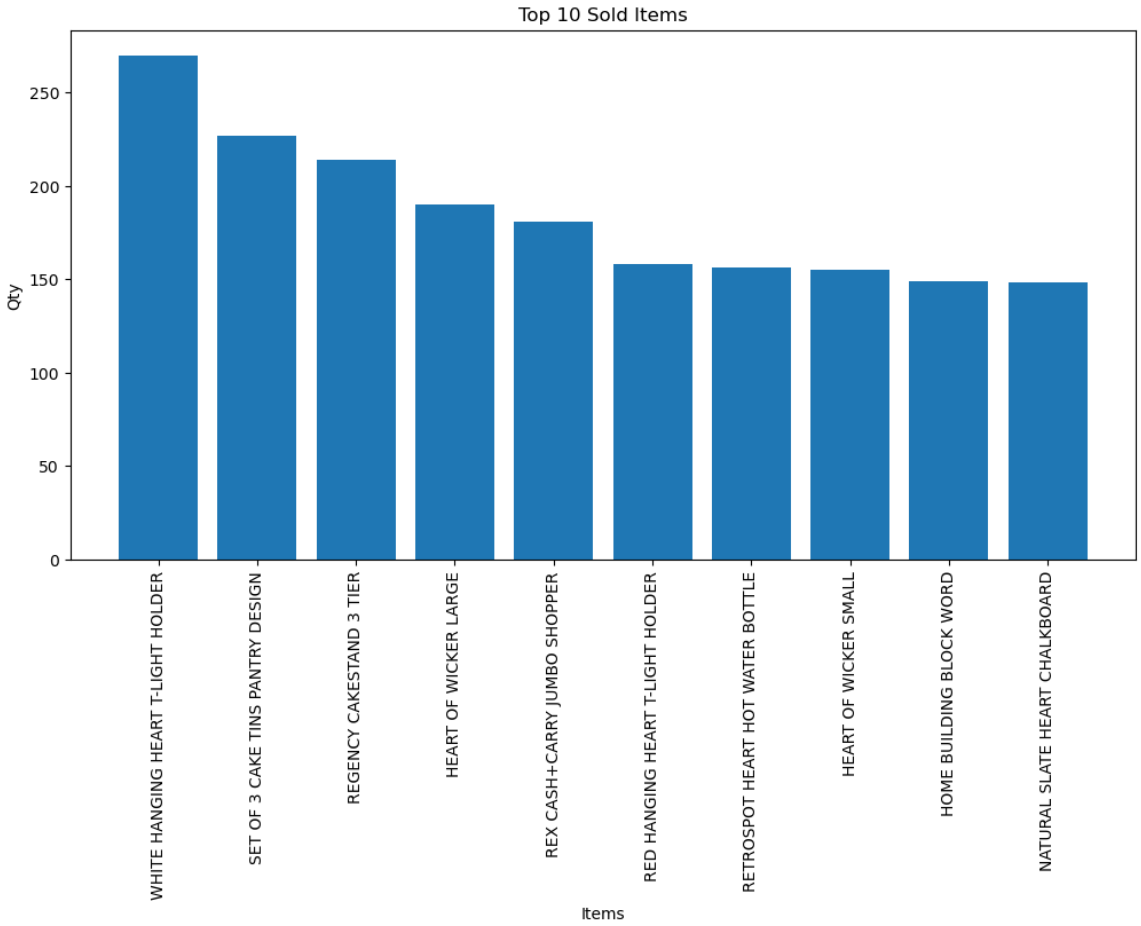

Figure 8: Top 10 products being purchased

4. Implementation and Testing

4.1. Recommendations Systems

The recommendations systems refer to predicting user preferences for any particular items (Leskovec, Rajaraman & Ullman 2014, p.287). The recommendations systems can be classified into a number of categories as shown in figure 9, however, we will discuss two broad type of recommendation systems including content based and collaborative filtering.

Content based Filtering: Content based filtering recommend items based on the interests of one user (Leskovec, Rajaraman & Ullman 2014, p.287). For instance, if a user buys several full sleeve shirts from Amazon, then the content based recommendation systems would suggest the user more full sleeve shirts from other suppliers.

Collaborative filtering: Collaborative filtering recommends items based on past user-item interactions in the form of a ratings matrix. It uses the assumption that users will have identical interests with similar behaviour. Hence, ratings from all other users are considered to recommend items to one who has similar interests (Kotu & Deshpande 2019, p.351).

This can be classified as two sub classes which include neighbourhood methodology and latent factor models. We will discuss the neighbourhood methodology as this has been used in our project. The neighbourhood methodology can be classified into two methods which are user based and item based. The user based methodology finds users with similar interests where their ratings are used to predict a new rating for a new user-item pair. On the other hand, item based methodology finds similar items and their ratings to predict new user-item pair (Kotu & Deshpande 2019, p.352).

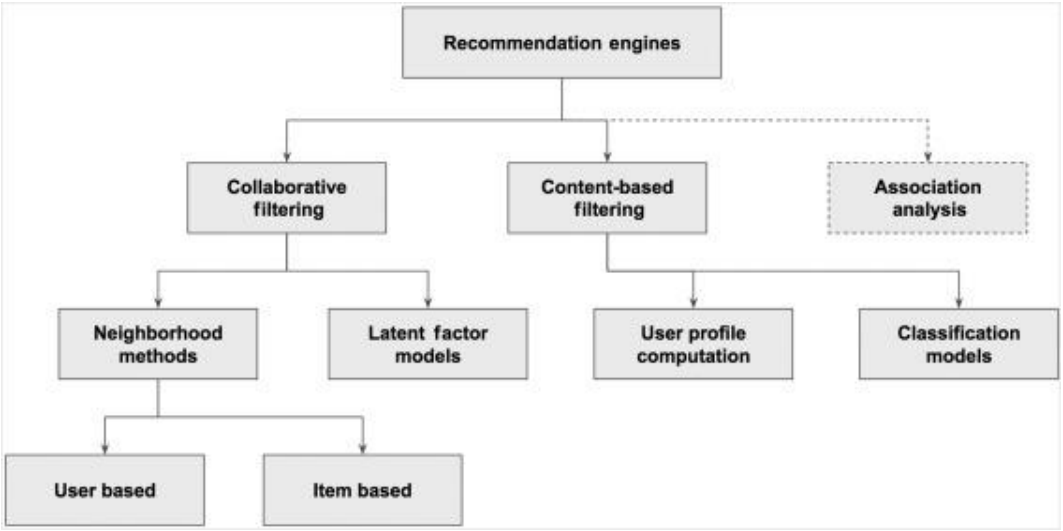

Figure 9. Classification of recommendation systems (Kotu & Deshpande 2019, p.351)

4.2. Training, Testing and Evaluation Methodology

The training phase involves fitting the model with training dataset. During training, the rocommendations systems learns from historical data, adjusts parameters to minimise errors in prediction. In our project the method for generating recommendations we have used item-item similarities using cosine similarity. First of all, we created user-item matrix to obtain the user-item pair and converted to sparse matrix format for efficient calculations. Secondly, we have transposed the sparse matrix so that it can be used for calculating cosine similarity among items and then converted into dataframe to generate recommendations. The model takes Customer ID as input to calculate similarity scores and provide top items with highest scores for the corresponding user.

Based on the training data here are the recommended items (Table 1) for our top 5 users.

Customer ID	Recommended Items
17841	'EDWARDIAN PARASOL NATURAL', 'SET OF 3 HEART COOKIE CUTTERS', 'HOME BUILDING BLOCK WORD', 'LUNCH BAG SPACEBOY DESIGN', 'LUNCH BAG BLACK SKULL.'
15311	'WHITE HANGING HEART T-LIGHT HOLDER', 'JUMBO BAG RED RETROSPOT', 'SET OF 72 PINK HEART PAPER DOILIES', 'HEART OF WICKER SMALL', 'SET OF 3 CAKE TINS PANTRY DESIGN'
12748	'SET OF 3 HEART COOKIE CUTTERS', 'PACK OF 72 RETROSPOT CAKE CASES', 'HAND OVER THE CHOCOLATE SIGN', 'RED HANGING HEART T-LIGHT HOLDER', 'BLUE SPOT CERAMIC DRAWER KNOB'
17850	'EDWARDIAN PARASOL NATURAL', 'WOOD BLACK BOARD ANT WHITE FINISH', 'ENGLISH

	ROSE HOT WATER BOTTLE', 'RETROSPOT HEART HOT WATER BOTTLE', 'CHOCOLATE HOT WATER BOTTLE'
15039	'SET OF 3 HEART COOKIE CUTTERS', 'SET/10 RED POLKADOT PARTY CANDLES', 'VICTORIAN SEWING BOX LARGE', 'WOOD 2 DRAWER CABINET WHITE FINISH', 'STRAWBERRY LUNCH BOX WITH CUTLERY'

Table 1: Recommended items for top 5 customers

We have also generated an output (Table 2) for top 5 items that the store may suggest to users without user IDs having no past purchase history. For instance, if someone is purchasing our top recommended item "WHITE HANGING HEART T-LIGHT HOLDER", the store can suggest the below items ranked according to similarity score.

	Itemname	Similarity
0	WOODEN PICTURE FRAME WHITE FINISH	0.691606
1	HAND WARMER UNION JACK	0.679674
2	GLASS STAR FROSTED T-LIGHT HOLDER	0.679450
3	RED WOOLLY HOTTIE WHITE HEART.	0.674411
4	KNITTED UNION FLAG HOT WATER BOTTLE	0.671572

Table 2: Recommended items for white hanging heart t-light holder

In the testing phase, we have fitted the test dataset to test our recommendations systems. It was done with user-item matrix from the test dataset. The system takes the customer ids from the test dataset to check if the user id is in the train dataset. If the test user id is found in the train dataset it will proceed with predicting successful top N (in our case it was 5) matches with the recommended items.

In the evaluation phase, we took actual purchased items and intersected with the predicted items for the successful purchase according to the recommendation. In order to evaluate the system we used the following metrics.

User coverage: Percentage of users who received recommendations.

Root-mean-square error (RMSE): The difference between predicted and real user-item rating.

Average Precision: Percentage of relevant recommendations which is the intersection of predicted and actual items.

Average recall: Percentage of relevant items that were recommended.

The following diagram (Figure 10) summaries the whole process of the system.

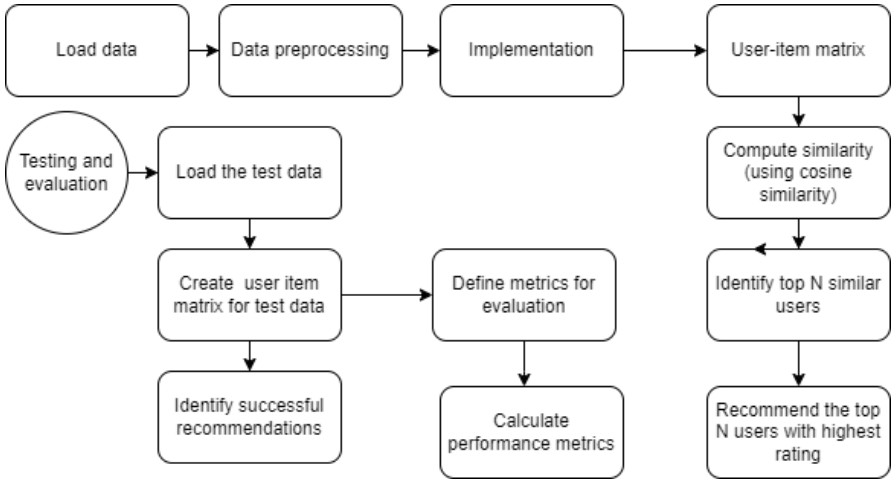

Figure 10: Diagram of the implemented recommendations systems

5. Discussion of Results

From our recommendations systems we wanted to know if our recommendations worked successfully. The results (Table 3) suggest that customers are purchasing according to the recommendations we made based on the similarity score which suggests a successful recommendation system.

Customer ID	Recommended Items	Purchased Items
12423	'RED RETROSPOT CHARLOTTE BAG', 'WOODLAND CHARLOTTE BAG', 'CHARLOTTE BAG SUKI DESIGN', 'CHARLOTTE BAG PINK POLKADOT', 'LOVE BUILDING BLOCK WORD'	'RED RETROSPOT CHARLOTTE BAG'
12528	'CERAMIC CAKE STAND + HANGING CAKES', 'SET OF 3 CAKE TINS PANTRY DESIGN', 'SET OF 4 ENGLISH ROSE PLACEMATS', 'RECIPE BOX PANTRY YELLOW DESIGN', 'APPLE BATH SPONGE'	'SET OF 3 CAKE TI NS PANTRY DESIG N'
12553	"POPPY'S PLAYHOUSE LIVINGROOM", 'RED 3 PIECE RETROSPOT CUTLERY SET', 'STRAWBERRY LUNCH BOX WITH CUTLERY', 'GREEN 3 PIECE POLKADOT CUTLERY SET', 'JUMBO BAG STRAWBERRY'	'STRAWBERRY LU NCH BOX WITH CU TLERY'
12728	"POPPY'S PLAYHOUSE KITCHEN", "POPPY'S PLAYHOUSE BEDROOM", 'OFFICE MUG WARMER POLKADOT', "POPPY'S PLAYHOUSE BATHROOM", 'COOKING SET RETROSPOT'	"POPPY'S PLAYHO USE KITCHEN", "P OPPY'S PLAYHOUS E BATHROOM"
12748	'SET OF 3 HEART COOKIE CUTTERS', 'PACK OF 72 RETROSPOT CAKE CASES', 'HAND OVER THE CHOCOLATE SIGN', 'RED HANGING HEART T-LIGHT HOLDER', 'BLUE SPOT CERAMIC DRAWER KNOB'	'HAND OVER THE CHOCOLATE SIG N'

Table 3: Recommended items vs purchased items

As a part of evaluation of the recommendation systems we have the metrics calculated as shown in table 4.

	Metrics	Value
0	User Coverage	0.427746
1	Average Precision	0.080180
2	Average Recall	0.029545
3	RMSE	0.990095

Table 4: Metrics for evaluation

The user coverage suggests that 42.7% of customers were recommended. Average precision and recall suggest 8.02% of the recommendations were relevant to the customers and 2.96% of recall suggests that 2.96% of all relevant items were successfully recommended to customers. The root mean square error of our recommendations systems 0.99 suggests that, on average, the predicted ratings for recommended items deviated from the actual ratings by approximately 0.99.

6. Conclusion and Recommendations

The project has successfully implemented a recommendation systems in order to improve sales both in store and online. The system provides effective recommendations to the customers based on their purchasing behaviour using collaborative filtering method.

It is crucial that the system efficiently handles a big volume of data. Our system is trained to generate recommendations for up to approximately 1 million customers which suggest the system's sustainability with the growing customer base. This provides the store to expand business in future with greater sales. Therefore, we recommend collaborative filtering method for the end users. However, there are room for improvements. In order to improve accuracy of the system the algorithm needs to be continuously continuous monitoring and performance evaluation using advanced machine learning techniques. Additional users data, for instance, browsing history, can be fed into the model training for better prediction. Mining frequent itemsets using apriori or fp-growth algorithms is also recommended to see if this can improve the prediction. By implementing these recommendations and continuously iterating on the recommendation system the store can personalise recommendations, improve customer satisfaction and increase sales.

7. Reflection

From the project I have learned how to develop a recommendations systems using a particular filtering method. I have used collaborative filtering methods in this case. However, in future, I would like to test the model using content-based filtering methods and compare both methods to determine the best possible recommendation systems. In addition, I would like to generate a system to mine frequent itemsets to see which method gives the more accurate recommendations.

8. References

Kotu, V & Deshpande, B 2019, *Data science: concepts and practice*, Morgan Kaufmann Publishers, Cambridge, MA, pp. 351-352, DOI:10.1016/C2017-0-02113-4.

Leskovec, J, Rajaraman, A & Ullman, JD 2014, *Mining of Massive Datasets*, Cambridge University Press, Cambridge, MA, p. 287.