A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Machine Learning/Data Visualization |
| *Evaluation Title:* | CA 2 |
| *Assessment Due Date:* | 24/11/2024. |
| *Presentation Date* | 29/11/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

[Womens Clothing E-Commerce Reviews 3](#_Toc183815008)

[1.Introduction 3](#_Toc183815009)

[1.1 Overview of the Project 3](#_Toc183815010)

[1.2 Objectives of the Analysis 3](#_Toc183815011)

[1.3 Significance of Machine Learning in E-Commerce 3](#_Toc183815012)

[1.4 Structure of the Report 4](#_Toc183815013)

[2. Data Preparation 4](#_Toc183815014)

[2.1 Loading and Inspecting the Dataset 4](#_Toc183815015)

[2.2 Removing Unnecessary Columns 5](#_Toc183815016)

[2.3 Handling Missing Values 5](#_Toc183815017)

[3. Recommendation System 5](#_Toc183815018)

[3.1 Introduction to Recommendation Systems 5](#_Toc183815019)

[3.2 Collaborative Filtering Implementation 5](#_Toc183815020)

[3.3 Insights and Results 7](#_Toc183815021)

[4. Market Basket Analysis 7](#_Toc183815022)

[4.1 Introduction to Market Basket Analysis 7](#_Toc183815023)

[4.2 Apriori Algorithm 8](#_Toc183815024)

[4.3 FP Growth Algorithm 9](#_Toc183815025)

[4.4 Comparative Analysis of Apriori and FP Growth 10](#_Toc183815026)

[5. Interactive Dashboard 11](#_Toc183815027)

[5.1 Distribution of Ratings by Age (18–35) 11](#_Toc183815028)

[5.2 Positive Feedback by Division (Age 18–35) 12](#_Toc183815029)

[5.3 Recommended vs Non-Recommended Products by Age (18–35) 13](#_Toc183815030)

[6. Discussion and Conclusion 13](#_Toc183815031)

[6.1 Summary of Key Findings 13](#_Toc183815032)

[6.2 Implications for E-Commerce Businesses 14](#_Toc183815033)

[6.3 Limitations and Recommendations for Future Work 15](#_Toc183815034)

[Conclusion 15](#_Toc183815035)

[Github Link 15](#_Toc183815036)

# Womens Clothing E-Commerce Reviews

## 1.Introduction

### 1.1 Overview of the Project

This report addresses the application of machine learning techniques in the field of online retailing. The project aims to increase customer satisfaction and develop business strategies by performing analyses on a dataset for women's clothing e-commerce. The study has been implemented in three main areas: recommendation systems, market basket analysis, and interactive dashboards.

In this context, the application of models such as collaborative filtering and frequent pattern analysis has been addressed to create data-driven strategies. The project makes significant contributions to the process of analyzing and optimizing customer behavior for e-commerce businesses.

### 1.2 Objectives of the Analysis

The main objectives of the study are as follows:

• Recommendation Systems: To provide personalized recommendations by analyzing users' purchase history and product preferences.

• Market Basket Analysis: To develop cross-selling strategies by identifying products that are likely to be purchased together.

These objectives aim to help e-commerce businesses respond to customer needs more quickly and effectively.

### 1.3 Significance of Machine Learning in E-Commerce

Machine learning plays a critical role in the e-commerce sector today. Its main advantages are as follows:

• Personalized Recommendations: Increases customer satisfaction by providing recommendations based on users' tastes and preferences.

• Optimized Inventory Management: Provides effective management of inventory by identifying frequently purchased products.

These technologies are used to improve customers' shopping experience and increase businesses' revenues.

### 1.4 Structure of the Report

This report consists of the following sections:

1. Introduction: A general summary of the project, its objectives and importance.

2. Data Preparation: Data cleaning, transformation and processing steps performed on the dataset.

3. Recommendation System: Application of user and product-based collaborative filtering methods.

4. Market Basket Analysis: Application and comparison of Apriori and FP Growth algorithms.

5. Interactive Dashboard: Visualization and analysis for users aged 18-35.

6. Discussion and Conclusion: Key findings, implications for business strategies, and recommendations for future work.

## 2. Data Preparation

### 2.1 Loading and Inspecting the Dataset

The dataset is the basis of the analysis. In this step, the dataset is loaded and its basic structure is examined. The dataset contains 23,486 rows and 11 columns related to users, products and feedbacks regarding women's clothing e-commerce.

Dataset columns:

• Clothing ID: Unique identification number for each product.

• Age: Customer's age.

• Rating: Customer evaluation scores (1–5).

• Recommended IND: Binary variable indicating whether the customer recommends the product.

• Positive Feedback Count: Number of positive feedbacks related to the product.

• Division Name, Department Name, Class Name: Indicates the categorical features of the product.

### 2.2 Removing Unnecessary Columns

• Columns that will not be used in the analysis on the dataset have been removed. For example:

• "Unnamed: 0" column contains only the ranking numbers.

• The "Title" and "Review Text" columns are text-based and are not used in the analysis in this project.

• Removing these columns made the data more workable.

### 2.3 Handling Missing Values

Missing values ​​were handled in a special way because they could affect the accuracy of the analysis:

• Categorical Columns: Missing values ​​in the "Division Name", "Department Name", and "Class Name" columns were filled with "Unknown". This eliminates the deficiency while maintaining the categorical consistency of the data.

• Numerical Columns: Since there is no missing numerical data, no processing was done in these fields.

## 3. Recommendation System

### 3.1 Introduction to Recommendation Systems

In this project, collaborative filtering methods that analyze users' past data and product preferences and provide personalized recommendations were implemented. Two basic approaches were examined:

1. User-User Collaborative Filtering: Based on similarities between users.

2. Item-Item Collaborative Filtering: Based on similarities between products.

These methods are critical for determining products that users may be interested in and increasing cross-selling opportunities.

### 3.2 Collaborative Filtering Implementation

#### 3.2.1 User-User Collaborative Filtering

User-User Collaborative Filtering analyzes similarities based on the scores users give to products. Similarities between users were calculated with the cosine similarity method and the most similar users were determined for the target user.

Findings:

• The most similar users for the 25-year-old target user are the 63, 42, 52, and 54 age groups, respectively.

• Recommendations were created for the target user based on the products that these users rated highly.

#### 3.2.2 Generating User-User Recommendations

Products that the target user had not previously evaluated, and that similar users rated highly, were recommended.

Recommended Products:

1. Clothing ID 863 (Estimated Score: 3.76)

2. Clothing ID 1033 (Estimated Score: 3.39)

3. Clothing ID 861 (Estimated Score: 3.27)

4. Clothing ID 1025 (Estimated Score: 3.19)

5. Clothing ID 975 (Estimated Score: 3.18)

These recommendations are products that are in line with the potential interests of the target user.

#### 3.2.3 Item-Item Collaborative Filtering

Item-Item Collaborative Filtering analyzes the similarities between products and recommends products that are similar to products that users have rated highly in the past. This method is especially advantageous for new users.

Findings Obtained:

• The similarity score between products was calculated based on the products that the target user rated highly.

• Recommendations were presented to the target user based on similar products.

#### 3.2.4 Generating Recommendations Based on Item Similarities

Products that are similar to products that the target user rated highly in the past were determined and recommended.

Recommended Products:

1. Clothing ID 873 (Similarity Score: 66.65)

2. Clothing ID 1083 (Similarity Score: 66.45)

3. Clothing ID 829 (Similarity Score: 65.62)

4. Clothing ID 862 (Similarity Score: 65.56)

5. Clothing ID 1022 (Similarity Score: 65.25)

These recommendations present products that may interest the target user based on similarities between the products.

### 3.3 Insights and Results

User-User Collaborative Filtering:

This method provides personalized recommendations by analyzing behavioral similarities between users. The recommendations created for the target user are based on products that similar users rate highly. Although the recommendations are personalized, their applicability to new users is limited.

Item-Item Collaborative Filtering:

This method provides more comprehensive recommendations by focusing on similarities between products. It is an effective method especially for new users. However, the level of personalization of the recommendations is lower compared to User-User Collaborative Filtering.

General Comment:

• User-User Collaborative Filtering: Provides more specific and personalized recommendations.

• Item-Item Collaborative Filtering: Provides broader recommendations and is more suitable for new users.

## 4. Market Basket Analysis

### 4.1 Introduction to Market Basket Analysis

Market Basket Analysis (MBA) is a method that aims to find meaningful relationships between products purchased together by customers in their shopping carts. This method plays a critical role especially in the development of cross-selling strategies and product recommendation systems.

Objectives:

• Understand which products customers tend to purchase together.

• Reveal hidden relationships between products.

• Optimize marketing and inventory management strategies with the relationships obtained.

In this analysis, two basic algorithms are used to extract relationship rules between products:

1. Apriori Algorithm

2. FP Growth Algorithm

These methods are used to model customers' purchasing behavior by identifying frequent item clusters and rules derived from these clusters. The analysis results provide important insights that will add value to e-commerce businesses.

### 4.2 Apriori Algorithm

#### 4.2.1 Implementation and Results

Apriori Algorithm is a basic data mining algorithm used to extract frequent itemsets and relationship rules based on these sets. This algorithm works according to minimum support and confidence thresholds.

Implementation of the project:

• Minimum support value is determined as 2%.

• Minimum confidence value is used as 100%, which shows that all proposed rules are reliable.

• In order to make the code work faster and to evaluate the algorithm efficiently, only the first 100 rows and 10 columns of the dataset were processed. This sample is sufficient to analyze the performance of the algorithm and see the basic relationships.

Findings: According to the data obtained from the images, the relationship rules derived by the algorithm are as follows:

1. (0) → (7):

o Support: 0.0476

o Confidence: 1.0

o Lift: 21.0

This result shows that when product 0 is purchased, the probability of purchasing product 7 is very high.

2. (4) → (3):

o Support: 0.0476

o Confidence: 1.0

o Lift: 21.0

This relationship shows that when product 4 is purchased, product 3 is almost always purchased.

Analysis Comments:

• Support: The fact that the rules are realized at a rate of 4.76% indicates that the rules are repeated frequently.

• Confidence: A confidence level of 100% indicates that the relationships are quite strong.

• Lift: A high lift value of 21 indicates that the probability of purchasing these product pairs together is much higher than randomness.

These results reveal that certain product pairs are frequently purchased together among customers. These insights can be used to develop cross-selling strategies.

### 4.3 FP Growth Algorithm

#### 4.3.1 Implementation and Results

FP Growth Algorithm is a fast and effective algorithm used to identify frequent itemsets. This algorithm works faster than Apriori Algorithm on large datasets by minimizing the need for data scanning.

Implementation of the Project:

• Only the first 100 rows and 10 columns of the dataset were used to make the code run faster and to evaluate the performance of the algorithm. This sample contains the same sample as Apriori, making it easier to make comparisons.

• The minimum support value was determined as 2%.

• The algorithm worked with the same goal as Apriori: to identify frequent itemsets and derive relationships based on these sets.

Findings Obtained: According to the data obtained from the visuals, the FP Growth Algorithm results are completely consistent with Apriori Algorithm. The relationship rules extracted are:

1. (0) → (7):

o Support: 0.0476

o Confidence: 1.0

o Lift: 21.0

This shows that if product 0 is purchased, product 7 is also likely to be purchased.

2. (4) → (3):

o Support: 0.0476

o Confidence: 1.0

o Lift: 21.0

A customer who purchased product 4 usually purchased product 3 as well.

Analysis Comments:

• Support: Both algorithms showed 4.76% support for frequent itemsets.

• Confidence: 100% confidence demonstrates the validity of the rules.

• Lift: High lift values ​​of 21 prove that product pairs are strongly related.

Advantage of FP Growth Algorithm:

• It required less processing time and provided the same results more efficiently.

• It proved to be a more effective algorithm compared to Apriori in large datasets.

### 4.4 Comparative Analysis of Apriori and FP Growth

#### 4.4.1 Similarities and Differences

Similarities:

• Both algorithms extracted the same rules:

o (0) → (7): Support: 0.0476, Confidence: 1.0, Lift: 21.0

o (4) → (3): Support: 0.0476, Confidence: 1.0, Lift: 21.0

• Support (%4.76), confidence (%100) and lift (21.0) values ​​are the same for both algorithms.

Algorithms were run on the same data sample (first 100 rows and 10 columns), which increased the comparability of the results.

Difference:

1. Processing Time:

o Apriori Algorithm: It required longer processing time. This may be a disadvantage in large datasets.

o FP Growth Algorithm: It provided the same results with less processing time. This algorithm is more advantageous in large datasets.

2. Data Scanning:

o Apriori performs multiple scans on the data to identify frequent item clusters.

o FP Growth works faster by minimizing the need for data scanning.

#### 4.4.2 Conceptual Insights

The rules derived by both algorithms clearly reveal the relationship between products. For example:

• Product 0 and Product 7: It is one of the product pairs that have a high tendency to be purchased together. This shows that cross-selling offers for these products can be effective.

• Product 4 and Product 3: Related products can be considered categorical or complementary. This situation can be included in product recommendation systems.

## 5. Interactive Dashboard

### 5.1 Distribution of Ratings by Age (18–35)

A graph with different colored bars

Description automatically generated

This visual shows the distribution of ratings given to products by users between the ages of 18–35. The Product Ratings axis shows the ratings given by users (from 1 to 5), and the count axis shows the number of times these ratings were given.

Analysis and Comment:

• General Satisfaction Trend:

The vast majority of users gave 4 and 5 stars. This shows that the overall satisfaction level is quite high and the products largely meet the expectations of the target audience.

• Low Rating Rates:

The fact that 1 and 2 star ratings are quite low reflects that the product quality is generally satisfactory. However, analyzing the reasons for these low ratings can support future improvement efforts.

• Medium Rating (3 Stars):

3 star ratings may indicate that the products do not fully meet the expectations of some users. Analyzing this group may provide opportunities to improve product quality.

### 5.2 Positive Feedback by Division (Age 18–35)

A graph of different colored lines

Description automatically generated with medium confidence

This image shows the distribution of positive feedback (Positive Feedback Count) given to products by users in the 18–35 age group according to the "Division" categories to which the products are affiliated. The Division Name axis represents the product categories, while the Feedback Count axis represents the total number of positive feedback given to each category.

Analysis and Interpretation:

• Intense Feedback: General Division received much more positive feedback than other categories. This shows that the products meet general customer needs and increase customer satisfaction.

• Low Feedback: Low feedback numbers in categories such as Intimates and Unknown may indicate that interest in these categories is limited or that the products in these categories do not meet customer expectations.

### 5.3 Recommended vs Non-Recommended Products by Age (18–35)

A graph of growth in a bar

Description automatically generated with medium confidence

This image examines the product recommendation rates of users in the 18–35 age group according to age range. The Age axis shows the age groups, and the Recommended IND axis shows the recommendation rates of the products. The color scale of the visual more clearly expresses the probability of products being recommended.

Analysis and Comment:

• General Satisfaction and Trust:

Increasing product recommendation rates may indicate that customer satisfaction increases with age. Customers' trust or satisfaction with products may become more apparent as age increases.

• Opportunities for the 18–25 Age Group:

The relatively low recommendation rates in younger age groups may indicate that strategies targeting this group need to be strengthened.

## 6. Discussion and Conclusion

### 6.1 Summary of Key Findings

In this study, various analyses were performed on women's clothing e-commerce data and the findings were discussed in detail. The main findings can be summarized as follows:

1. Data Cleaning and Preparation:

o Missing and unnecessary columns were cleaned, categorical missing values ​​were filled with "Unknown" to make the data suitable for analysis.

o The data structure was analyzed with sufficient variety and scope to understand customer behavior.

2. Recommendation Systems:

o User-User Collaborative Filtering and Item-Item Collaborative Filtering methods were used.

o User-User Filtering provided personalized recommendations, while Item-Item Filtering provided recommendations to increase new product discovery.

o For example, product recommendations for users aged 25 were derived based on both user similarities and product similarities.

3. Market Basket Analysis:

o Apriori and FP Growth algorithms were applied, and both algorithms revealed frequently purchased product pairs.

o High lift values ​​(e.g. 21) showed that the tendency of products to be purchased together is much higher than random.

4. Interactive Dashboard:

o Visualization studies provided important insights into rating behaviors, product recommendation rates, and popular product categories in different age groups.

o For example, the popularity of the Dresses category and the strong feedback rates of the General Division stood out.

### 6.2 Implications for E-Commerce Businesses

These analyses have many useful results for e-commerce businesses:

1. Personalized Experiences:

o User-based recommendation systems can increase customer satisfaction and encourage repeat purchases.

o Personalized campaigns can be more effective for younger age groups.

2. Product Strategies:

o Investments in popular categories (e.g. Dresses, Intimates) can be increased.

o Feedback analyses can be conducted for less popular categories and improvement opportunities can be evaluated.

3. Cross-Selling Opportunities:

o Based on the Market Basket Analysis results, cross-selling strategies can be implemented for product pairs with a high tendency to be purchased together.

### 6.3 Limitations and Recommendations for Future Work

Limitations of this study and recommendations for future work:

1. Limited Data Sample:

o Some analyses were performed by reducing the sample size (e.g., first 100 rows and 10 columns). Broader insights can be obtained when working with larger datasets.

2. Incomplete Feedback:

o Low feedback rates in categories such as Unknown Division should be supported with more information.

## Conclusion

This project has provided a comprehensive analysis to support data-driven decision making in the women's clothing e-commerce sector. Approaches such as Recommendation Systems, Market Basket Analysis, and Interactive Dashboard provide important tools to increase customer satisfaction and optimize sales. In the future, these analyses can be expanded with larger datasets and new features.

## Github Link

https://github.com/TaylanOzgur96/MachineLearning-DataVisualizationCA2