Products Recommendation

WÜRTH ITALIA – Capstone Projects

Gyeongwon Song

June 19, 2024

1 Introduction

This project aims to develop an effective recommendation system for Würth Italia using user interaction data. By analyzing both online and offline customer behaviors, the project seeks to increase customer satisfaction and boost sales through accurate, personalized recommendations.

2 Missing values

Inside the data frame, in the 'dat_action' column, there are 6379695 null values, and in the 'actions_type' column, there are 6379695 instances of 'no_data'. Both missing values are connected. Given this connection, to ensure data quality and avoid potential biases in our analysis, we decided to remove all of these missing values.

3 Exploratory Data Analysis

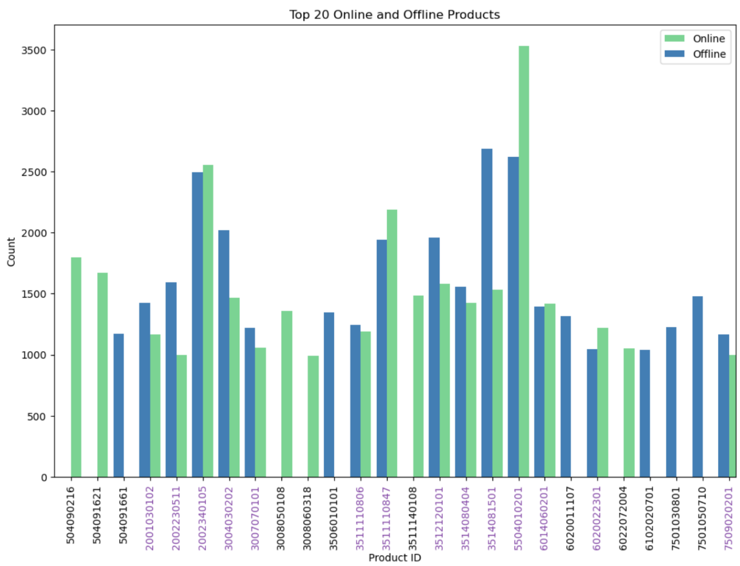

Figure 1: The most active products top-20 based on online activity

The overlap of 14 out of 20 top-selling products across both online and offline channels underscores strong consumer demand and consistent preferences.

3.1 Data Sparsity

The data sparsity observed in both user and item rating distributions can pose significant challenges for collaborative filtering algorithms, which rely on sufficient overlapping user-item interactions to make accurate recommendations. Sparse data can lead to less reliable and less personalized recommendations due to the lack of sufficient information.

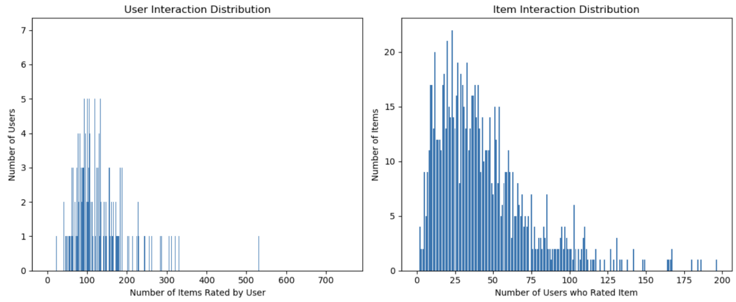

Figure 2: Distribution of User and Item Ratings: Indicators of Data Sparsity

4 Products Recommendation

4.1 Customer-Item Matrix

level5id	46228	501010204	501010206	501090140	503010409	503010517	503010518	503011304	503011305	503011306	...	50011201539	60150111220	6(
customerid														
11448266	0	0	0	0	0	0	0	0	0	0	...	0	0	
30874266	0	0	1	0	0	0	0	0	0	0	...	0	0	
50597266	0	0	0	0	0	0	0	0	0	0	...	0	0	
55173266	0	1	1	0	0	0	0	0	0	0	...	0	0	
63016266	0	0	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 942 columns

First, we calculate the cosine similarity between users based on their interactions with items. Next, we set up the similarity matrix with user IDs as both columns and index.

4.2 User-Based Collaborative Filtering

Recommendation Analysis: User A vs. User B

1. Computed cosine similarity between users to measure their similarity based on interactions.
2. Identified items that User A purchased but User B did not, based on their purchase histories.
3. This analysis yielded a set of 75 items that can be recommended to User B, leveraging User A's preferences.

4.3 K-Nearest Neighbors (KNN) Algorithm

Finds the nearest neighbors of a user based on their interaction data and recommends products that these neighbors have interacted with.

KNN Implementation

1. Normalized the user-item matrix to ensure consistency in the data range.
2. Initialized the KNN model using cosine distance and employed the brute-force algorithm for computation.
3. Developed a function to generate recommendations for a user (e.g., '9935934266') based on the fitted KNN model.

5 Conclusion

The project developed a recommendation system using collaborative filtering and KNN algorithms. The Product Recommendation identifies similarities between users and suggests products based on their past interactions. The analysis of dataset sparsity highlighted the challenges of working with sparse data and the importance of advanced techniques to mitigate these challenges.