Introduction

In the current ecommerce landscape it is necessary to provide engaging user experiences for better customer retention. Most companies like Amazon have a section dedicated to product recommendations that showcase products that closely match the current product the user is viewing. There are more chances that the customer will likely purchase through the business website if they find a product closely resembling what they are searching for or already looking at. With this project, my goal was to make a sentiment-aware product recommendation tool or system through PySpark. There are 2 main models used in this project-

1. A sentiment classification model which predicts if the customer reviews for a product are positive or negative using logistic regression.
2. A recommendation model that accounts for product ratings, matching product categories and the textual similarity from customer reviews.

The dataset is related to sentiment based product recommendations and reviews and it was obtained from Kaggle. It contains over 30,000 product reviews across over 200 categories and over 20,000 users. Here is the link to the dataset page https://www.kaggle.com/datasets/marusagar/sentiment-based-product-recommendation-and-reviews/data.

The dataset combines numerical user ratings with user feedback about products and the overall user sentiment for a particular product. It also appropriately includes the category of the products purchased. This format of the data allows us to create a hybrid recommendation model.

Methodology

Step 1: Data Preparation

The filtered raw dataset contained columns with information about product names, user review texts, user numeric ratings, labelled sentiment and categories.
Data cleaning involved the following steps:

- Removing missing or null entries, and replacing missing user sentiment values using customer ratings.
- Formatting product categories through splitting and normalizing the comma-separated strings.
- Cleaning and lowercasing the user reviews text to remove punctuation and special symbols.

After data preprocessing, model setup to do sentiment classification and provide recommendations was possible

Step 2: Sentiment Classification

Using a TF-IDF pipeline to train a logistic regression classifier, each review was given a positive or negative prediction label. The model was evaluated using the confusion matrix analysis to get accuracy, precision, recall and F1 score based on the correctness of predicting the positive reviews. Thus, the model was able to classify the given customer review as positive or negative with a label of 0 for negative and 1 for positive.

Step 3: Positive Reviews filtering

Only the products with positive sentiment prediction based on the sentiment classification and those with a review rating higher or equal to 4 were considered going forward. For each of these selected products, the reviews were aggregated and their average rating and representative categories were also computed.

Step 4: TF-IDF Feature Representation

Each product's aggregated review was converted into a TF-IDF vector, which showed the importance of each word in relation to all products in the updated dataset. Similar products had similar high dimensional vector representations which were used to determine the cosine similarity between them.

Step 5: Hybrid Recommendation Model

The final recommendation engine took into account the cosine similarity between all the products' TF-IDF vectors and the average product rating. Then a hybrid final score was computed using the weighted sum of the cosine similarity and normalized rating, with more weight assigned to the similarity.
Another filter was added to consider only products with 2 matching categories with the target product to provide some more context to the recommendations. The recommendation system finally ranked all the candidate products by this hybrid scoring and gave the top 5 recommendations for each product.

Results

Sentiment classification evaluation results->

```
Logistic regression confusion matrix [[ 444.  210.]
 [ 334. 4996.]]
Corrected metrics
Accuracy: 0.9090909090909091
Precision (label 1): 0.9596619285439877
Recall (label 1): 0.9373358348968105
F1 Score (label 1): 0.9483675018982536
```

```
Test AUC: 0.8953475796225845
```

Product recommendations and evaluations for 5 sample products->

```
product Lundberg Organic Cinnamon Toast Rice Cakes
mean_similarity 0.246400133705376 42
mean_rating 4.818403779422217
```

name	categories	average_rating	similarity	final_score
Stacy's Garden Veggie Medley Pita Chips	Food,Packaged Foods,Snacks,Chips & Pretzels,Snacks, Cookies & Chips,Chips,Grocery & Gourmet Food,Snack Foods,Chips & Crisps,#10214 in,Seasonal,Grocery	4.6	0.616538715789092	0.7015771010523645
Chex Muddy Buddies Brownie Supreme Snack Mix	Food,Packaged Foods,Snacks,Snacks, Cookies & Chips,Chips,Home & Garden,Food & Beverages,Snack Foods,Other Snack Foods,grocery,Toys,ARCHIVE SHELF,CAT	4.761904761904762	0.24874407976583265	0.4598351415503684
Annie's Homegrown Gluten Free Double Chocolate Chip Granola Bars	Food,Packaged Foods,Snacks,Cereal Bars and Granola Bars,Food & Beverage,Cookies, Chips & Snacks,Granola Bars,Snacks, Cookies & Chips,Granola Bars & Snack Bars,Breakfast & Cereal,Breakfast Bars,Grocery & Gourmet Food,Snack Foods,Bars,Granola,Home & Garden,Food & Beverages,Snack Bars,Featured Brands,Grocery,General Mills,Granola & Nutrition Bars	5.0	0.19933812089444416	0.4395366846261109 5
Tostitos Simply Blue Corn Tortilla Chips	Food,Packaged Foods,Snacks,Chips & Pretzels,Snacks, Cookies & Chips,Chips,Food & Beverage,Cookies, Chips & Snacks,Food & Beverage Ways To Shop,Tailgating Essentials,Grocery & Gourmet Food,Snack Foods,Chips & Crisps,Tortilla	4.882352941176471	0.08690570191097743	0.3537751678082725
Chester's Cheese Flavored Puffcorn Snacks	Food,Packaged Foods,Snacks,Chips & Pretzels,Food & Beverage,Cookies, Chips & Snacks,Chips,Snacks, Cookies & Chips,Food & Beverage Ways To Shop,Tailgating Essentials,Grocery,#1,079 in,#236 in,#26 in	4.947761194029851	0.08047480816653568	0.3531975867583668 3

```
product Batherapy Natural Mineral Bath Sport Liquid, 16 oz
mean_similarity 0.05476897165252413
mean_rating 4.894511920827711
```

name	categories	average_rating	similarity	final_score
Jergens Extra Moisturizing Liquid Hand Wash, 7.5oz	Personal Care,Bath, Shower & Soap,Liquid Hand Soap,Hand Soap & Sanitizers,Beauty,Body Cleansers	5.0	0.06915002053267584	0.3484050143728726
Yes To Carrots Nourishing Body Wash	Personal Care,Bath, Shower & Soap,Body Wash & Cleanser,Bath & Body,Body Wash & Cleansers,Cleansers,Body Scrubs,Scrubs & Body Treatments,Bath & Body Care,Beauty,Body Washes	4.929824561403508 6	0.05388661689732258 6	0.3334541055123363 4
J.R. Watkins Hand Cream, Lemon Cream	Personal Care,Skin Care,Hand Cream,Beauty,Body Lotions & Creams,Natural Beauty,Natural Personal Care,Natural Skin Care,Ways To Shop,Bath & Body,Hand Creams & Lotions	5.0	0.04188038303521101	0.3293165812465051
Mill Creek Aloe Vera & Paba Lotion	Personal Care,Skin Care,Moisturizer,Bath & Body,Hand Creams & Lotions,Body Lotions,Beauty	4.888888888888889	0.04574697243869287	0.3253562140441837
Yes To Grapefruit Rejuvenating Body Wash	Personal Care,Bath, Shower & Soap,Body Wash & Cleanser,Bath & Body,Body Wash & Cleansers,Beauty,Bath & Body Care,Scrubs & Body Treatments,Body Scrubs	4.653846153846154	0.06326041804178086	0.3235136618600153

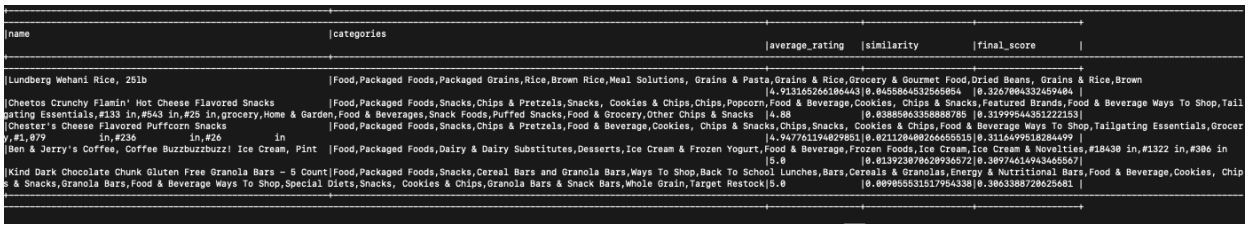
```
product Iman Luxury Moisturizing Lipstick, Black Brandy 006
mean_similarity 0.23799439473196732
mean_rating 4.828540229885058
```

name	categories	average_rating	similarity	final_score
Burt's Bees Lip Shimmer, Raisin	Personal Care,Makeup,Lipstick, Lip Gloss, & Lip Balm,Lip Gloss,Beauty,Lips,Beauty & Personal Care,Skin Care,Lip Care,Lip Balms & Treatments	4.851034482758621	0.3195174741562711	0.5147243008749071
L'oreal Paris Colour Caresse Wet Shine Stain, Pink Resistance	Personal Care,Makeup,Lipstick, Lip Gloss, & Lip Balm,Lip Gloss,Beauty,Lips	4.666666666666667	0.31774895696132555	0.5024242698729279
Burt's Bees Lip Shimmer, Watermelon	Personal Care,Makeup,Lipstick, Lip Gloss, & Lip Balm,Lip Balm,beauty,Lips,Lip Glosses,Cosmetics	4.625	0.2301543059856023	0.4386808139189923
Weleda Everon Lip Balm	Personal Care,Makeup,Lipstick, Lip Gloss, & Lip Balm,Lip Balm,Beauty,Lip Care	5.0	0.18882657586134532	0.4321786831029417 6
Soothing Touch Lemon Cardamom Vegan Lip Balm .25 Oz	Personal Care,Makeup,Lipstick, Lip Gloss, & Lip Balm,Lip Balm,Balms & Moisturizers,Beauty,Skin Care,Lip Care,Lips,Balms	5.0	0.13372466108233455	0.3936072675763425

```
product 2017-2018 Brownline174 Duraflex 14-Month Planner 8 1/2 X 11 Black
mean_similarity 0.020599125629830 27
mean_rating 4.865999775205125
```

name	categories	average_rating	similarity	final_score
Avery174 Ready Index Contemporary Table Of Contents Divider, 1-8, Multi, Letter	Office,Office Supplies,Office Organization,Binders and Accessories,All Binders,School & Office Supplies,Filing,File Dividers,Dividers	4.919860627177701	0.040347325105291 01	0.3234347652043657 6
Pendaflex174 Divide It Up File Folder, Multi Section, Letter, Assorted, 12/pack	School & Office Supplies,Filing,Files,File Folders,Office,Office Supplies,Office Organization,Filing and Folders,All Folders and Filing,Project Folders,Office Products,Filing Products,Folders,electronics,Filing & Storage,File Boxes	4.418138248847926	0.05793551409825512	0.3051631561996542
Smead174 Recycled Letter Size Manila File Backs W/prong Fasteners, 2 Capacity, 100/box	School & Office Supplies,Filing,Files,File Folders,Office,Small Business Center,Small Business Bulk Buys,All Bulk Buys,Office Supplies,Office Organization,Home	5.0	0.0035807648676036623	0.3025065354073226
Avery174 11-1/4 X 9-1/4 Index Maker Extra Wide Label Dividers With 5 Tab - Clear (5 Sets Per Pack)	School & Office Supplies,Filing,File Dividers,Labels & Label Makers,Sticker Labels,Office,Office Supplies,Office Organization,Binders and Accessories,All Binders,Ways To Shop,Classroom Essentials,Dividers	5.0	8.591348449624736E-4	0.3006013943917376
Smead174 2 1/4 Inch Accordion Expansion Wallet, Poly, Letter, Translucent Green	School & Office Supplies,Filing,Files,File Folders,Seasonal,Back to School Top Items,Office,Office Supplies,Office Organization	5.0	2.7288723303990854 6E-4	0.3001918210631274

```
product RC Cola, 12oz
mean_similarity 0.025707217850187935
mean_rating 4.948185292027259
```



name	categories	average_rating	similarity	final_score
Lundberg Wehani Rice, 25lb	Food,Packaged Foods,Packaged Grains,Rice,Brown Rice,Meal Solutions, Grains & Pasta,Grains & Rice,Grocery & Gourmet Food,Dried Beans, Grains & Rice,Brown	4.913652661064 43	0.045586453256 5054	0.326780433245 9404
Cheetos Crunchy Flamin' Hot Cheese Flavored Snacks	Food,Packaged Foods,Snacks,Chips & Pretzels,Snacks, Cookies & Chips,Chips,Popcorn,Food & Beverage,Cookies, Chips & Snacks,Featured Brands,Food & Beverage Ways To Shop,Tailgating Essentials,#133 in,#543 in,#25 in,grocery,Home & Garden,Food & Beverages,Snack Foods,Puffed Snacks,Food & Grocery,Other Chips & Snacks	4.88	0.0388506335888 8785	0.319995443512 22153
Chester's Cheese Flavored Puffcorn Snacks	Food,Packaged Foods,Snacks,Chips & Pretzels,Food & Beverage,Cookies, Chips & Snacks,Chips,Snacks, Cookies & Chips,Food & Beverage Ways To Shop,Tailgating Essentials,Grocery,#1,879 in,#236 in,#26 in	4.947761194029851	0.021120400266 6555515	0.311649951828 4499
Ben & Jerry's Coffee, Coffee Buzzbuzzbuzz! Ice Cream, Pint	Food,Packaged Foods,Dairy & Dairy Substitutes,Desserts,Ice Cream & Frozen Yogurt,Food & Beverage,Frozen Foods,Ice Cream,Ice Cream & Novelties,#18430 in,#1322 in,#306 in	5.0	0.013923070620 93657 2	0.309746149434 8556 7
Kind Dark Chocolate Chunk Gluten Free Granola Bars - 5 Count	Food,Packaged Foods,Snacks,Cereal Bars and Granola Bars,Ways To Shop,Back To School Lunches,Bars,Cereals & Granolas,Energy & Nutritional Bars,Food & Beverage,Cookies, Chips & Snacks,Granola Bars,Food & Beverage Ways To Shop,Special Diets,Snacks, Cookies & Chips,Granola Bars & Snack Bars,Whole Grain,Target Restock	5.0	0.009055531517 954338	0.306338872062 5681

Discussion

On running the recommendation system through 5 random products, the average cosine similarity ranged between 0.02 and 0.25, and the mean rating ranged from 4.8 to 4.9.So although the products being recommended are of high quality as we expected them to be through our filtration logic, the similarity between review texts is very low. The low cosine similarity values show that the user reviews may not share matching vocabulary across products even within similar categories. So,even though review similarity might be limited, the system was able to find highly rated products of matching categories, proving that the hybrid approach is still practical.

Conclusion

The project was successful in creating a sentiment-driven hybrid recommendation system through the combination of user ratings, user reviews, product categories and given plus predicted sentiment. We see high scores in all evaluation metrics for the logistic regression used to determine the sentiment label in the first modeling phase.

However, in the second phase of the project, due to the lack of any ground truth such as correct real recommendations to compare the obtained results to, an alternative evaluation approach was used. Using the mean cosine similarity and mean rating made the most sense to check the performance.

Also, due to the dataset only consisting of limited products and some reviews being too short or less descriptive or possibly too unique, it was hard to achieve high cosine similarities between products.

While the model was able to recommend most similar, highly rated products that belonged to matching categories, there can be further improvements made to it. Using context word embeddings like Word2Vec or BERT would be a great next step and having a larger, more diverse review set would help with better textual similarity.