**Personalized Apparel Recommendation In   
E-commerce**

-CONTENT BASED APPROACH

**Fakhera Mansuri J.**

Data Science

CC-406 Project-Ⅱ

# **CONTENTS**

introduction 2

oBJECTIVE 3

literature review 4

DATASET DESCRIPTION 5

METHODOLOGY 6

EXPECTED OUTCOMES 15

# **INTRODUCTION**

Our project, "Personalized Apparel Recommendation Engine for Small Businesses in E-commerce," aims to address the specific needs of small businesses lacking advanced intelligent systems like those used by industry giants such as Amazon. Small businesses often face challenges in providing personalized experiences due to resource and expertise constraints. By leveraging a content-based approach and the powerful Amazon API dataset, our recommendation engine will enable small businesses to cater to individual customer preferences and enhance their competitiveness in the e-commerce market. The Recommender System helps in suggesting the right content to the right user, thus helping in building a better user experience.

Our personalized apparel recommendation engine has the potential to significantly impact small businesses. By providing a functional and efficient system, it levels the playing field, enabling small businesses to compete with larger players in the market. Customers will benefit from a more personalized shopping experience, leading to increased satisfaction and loyalty. Moreover, our project aligns with broader goals of democratizing advanced technologies and supporting the growth of small businesses in the digital era. Through our tailored solution, we aim to make a positive impact on the e-commerce landscape for both small businesses and their customers.

We collected data for 1,83,000 products by using Amazon's Product Advertising API to gather the information in a way that complied with all applicable policies. We collected numerous features for each product, like the Title, Brand, Colour, Image-URL, Price, etc.

# **OBJECTIVE**

* The main objective of our project is to build a recommendation engine that caters to the needs of small businesses in the apparel industry and recommending similar apparel items to the user.
* Empowering Small Businesses:
  + Our recommendation engine will enable small businesses to compete effectively with larger players by offering personalized recommendations tailored to each customer's preferences.
* Enhancing Customer Shopping Experience:
  + By implementing our recommendation engine, small businesses can enhance their customers' shopping experience with personalized and curated apparel suggestions.
  + Customers will benefit from discovering new styles, finding items that align with their preferences, and saving time by having a curated selection of options.
* Increased Customer Engagement and Loyalty:
  + Personalized recommendations create a stronger connection between customers and small businesses, leading to increased engagement and brand loyalty.
  + The tailored suggestions will help customers feel understood and valued, driving repeat purchases and long-term relationships.
* Boosting Sales and Revenue:
  + The personalized recommendations generated by our engine will drive conversions and increase sales for small businesses.
  + By delivering targeted recommendations, small businesses can maximize their revenue potential and compete more effectively in the e-commerce market.

# **LITERATURE REVIEW**

Md Zaid Ahmed (2022) conducted a study presenting a system for e-commerce product recommendations using sentiment analysis and machine learning. This approach involves NLP for analyzing reviews, considering sentiments and ratings, and utilizing collaborative filtering for enhanced accuracy. The research highlights the applicability of the system using Amazon's platform.

Rohit Dwivedi (2022) investigated the application of recommendation systems to address the challenges of extensive e-commerce listings. The study explores diverse approaches, focusing on collaborative filtering and popularity-based engines using Amazon electronics data. By suggesting the top 5 products to users, these systems enhance user experience and potentially increase sales.

Mohammad R. Rezaei( 2021) explored using various models including a deep neural network to predict review ratings for music tracks on Amazon. With a dataset of 200,000 samples, the study aimed to enhance recommender systems for Amazon's products based on customer ratings and text reviews. The research highlights the effectiveness of the DNN approach alongside traditional models.

# **DATASET DESCRIPTION**

* The project leverages the Amazon API dataset, a comprehensive repository comprising details of approximately 180,000 products. These details encompass a range of vital attributes, including but not limited to product titles, brands, colors, images, and more.
* While the raw dataset includes 19 features for each product, our recommendation engine will focus on a subset of 6 key features, ensuring the generation of pertinent and insightful recommendations.

The selected **features** that our recommendation engine considers are as follows:

1. Amazon Standard Identification Number (ASIN): A unique identifier assigned to each product by Amazon, facilitating accurate tracking and referencing.
2. Brand: Denoting the brand associated with the product, this attribute contributes to a product's identity and consumer perception.
3. Color: This attribute encapsulates color information about the apparel, accommodating multiple color values such as "red and black stripes."
4. Product Type: Categorizing the type of apparel, such as "SHIRT" or "TSHIRT," providing insights into the product's style and intended use.
5. Medium Image URL: A link to the visual representation of the product, enabling image-based analysis and enhancing the recommendation process.
6. Title: The product's title, a concise textual representation containing valuable information about its attributes, style, and features.
7. formatted\_price :price of the product

# **METHODOLOGY**

Our project employs a multi-faceted approach to develop a personalized apparel recommendation engine for small businesses. The following steps outline our methodology:

1. **Data Acquisition**:

Utilize Amazon's API to access a diverse dataset of apparel products. Gather attributes such as product title, brand, description, category, price, user ratings, and reviews. This data forms the foundation for our recommendation engine.

1. **Data Cleaning**:

In this phase of our project, we meticulously address the quality and completeness of our dataset. Ensuring the accuracy and consistency of the data is a pivotal step towards reliable and meaningful results.

2.1 Handling Missing Data and Feature Insights

We begin by investigating missing data in various features to gain insights into our dataset.

* Product Type Name

- There are 72 unique product types.

- Shirts constitute 91.62% (167794 out of 183138) of the products.

* Brand

- We have 10577 distinct brands.

- There are 151 instances of missing brand information.

* Color

- We find 7380 distinct colors.

- Black is the most prevalent color (7.2%).

- Color information is available for approximately 35.4% (64956 out of 183138) of products.

* Formatted Price

- Only 15.5% (28395 out of 183138) of products have price information.

* Title

- All products are accompanied by descriptive and informative titles.

2.2 Duplicates Detection and Removal

Duplicates in the dataset can skew our analysis and recommendations. Our approach involves systematic identification and removal of duplicate items.

**Understanding Duplicates**-

- We identify 2325 products sharing the same title but differing in color.

* Duplicate Removal - Part 1

- We eliminate products with brief titles.

- The dataset is sorted based on titles in alphabetical order.

- Titles that are adjacent and closely similar are removed.

- This step significantly reduces data points from **183K to 17K.**

* Duplicate Removal - Part 2

- Further refinement is achieved by detecting non-adjacent but strikingly similar titles.

- After this step, our final cleaned dataset consists of **16K** data points.

By addressing missing data, gaining insights from features, and systematically removing duplicates, we ensure the integrity of our dataset, paving the way for accurate analysis and insightful results.

1. **Text Preprocessing:**

In this phase of the project, the textual attributes of the apparel products are processed to ensure their suitability for further analysis and recommendation.

* Data Retrieval and Initialization

The project starts by loading the apparel data from a previously pickled file ('16k\_apperal\_data') using Pandas. This dataset serves as the foundation for the entire project's development.

* Stopword Removal

NLTK's (Natural Language Toolkit) stop words are downloaded and utilized to create a set of common stop words in the English language. These stop words are essentially words that are common and generally considered to be uninformative, like "and", "the", "is", etc. The removal of stop words helps in reducing noise and focusing on more meaningful words.

* Text Pre-processing Function

A function named `nlp\_preprocessing` is defined to process each individual text (in this case, the product titles). This function takes three parameters: the text to be processed, its index in the dataset, and the column it belongs to.

* Text Cleaning

Within the function, the text is iterated through word by word. Special characters are removed using list comprehension, ensuring that only alphanumeric characters are retained. This step removes punctuation and other non-alphanumeric characters from the words.

* Lowercasing

All the words in the text are converted to lowercase to ensure consistency in comparison and analysis. This avoids treating the same word in different cases as different words.

* Stopword Removal (Continued)

The words are then checked against the list of stop words. If a word is not in the set of stop words, it's retained; otherwise, it's discarded. This process significantly reduces the number of unimportant words in the text.

* Text Transformation

The processed words are then concatenated to form a cleaned and preprocessed version of the text. This cleaned version replaces the original text in the dataset for the specific index and column.

* Iteration and Processing

The above `nlp\_preprocessing` function is applied to each product title in the dataset using an iteration over the dataset rows.

* Stemming (Optional - Not Highly Effective)

An attempt is made to apply stemming to the words in the product titles using the Porter stemming algorithm from NLTK. Stemming aims to reduce words to their base or root form. However, it's mentioned that this approach didn't work particularly well in this context.

* Execution Time Measurement
* The time taken for the preprocessing process is recorded using the `time.clock()`,the preprocessing of titles took approximately 3.57 seconds.
* Data Saving

Finally, the preprocessed dataset is saved as a pickled file ('16k\_apperal\_data\_preprocessed') for further use.

By performing these text preprocessing steps, the textual attributes of the apparel products are transformed into a format that's more amenable to subsequent analysis and recommendation techniques.

1. **Text-based Product Recommendation:**

4.1-Text-based Product Recommendation

We have implemented a text-based recommendation system that leverages the content of product titles to identify similar apparel items. This system involves various stages:

* Data Preparation

- We loaded a preprocessed dataset containing apparel information, including attributes such as ASIN, brand, color, medium image URL, product type name, title, and formatted price.

* Bag of Words (BoW) Model

- We constructed a Bag of Words (BoW) representation for the product titles using the `CountVectorizer` from the scikit-learn library.

- The BoW representation transforms each title into a numerical vector, where each element corresponds to the count of a specific word in the title.

- Cosine similarity was used as the distance metric to measure the similarity between the input product and other products in the dataset.

- For a given product, we computed the similarity scores and identified the top 'num\_results' products with the smallest distances.

It’s a first-cut solution

* TF-IDF Model

- We implemented a Term Frequency-Inverse Document Frequency (TF-IDF) model using the `TfidfVectorizer` from scikit-learn.

- TF-IDF assigns weights to words based on their frequency in a specific title and their rarity across the entire corpus.

- The TF-IDF vectors represent the titles, enabling comparison between product titles.

- Similar to the BoW model, we calculated cosine similarity to determine the similarity between products.

* Inverse Document Frequency (IDF) Model

- In this variation of the TF-IDF model, we incorporated the Inverse Document Frequency (IDF) value of words.

- We calculated IDF values for each word in the corpus and adjusted the TF-IDF vectors accordingly.

- This model emphasizes the importance of less frequent words in distinguishing between product titles.

4.2-Text Semantics based Product Similarity

Semantic Similarity or Semantic Textual Similarity is a task in the area of Natural Language Processing (NLP) that scores the relationship between texts or documents using a defined metric.

For eg. we know that the words 'tiger' and 'leopard' are related, also 'zebra' and 'strips' related.

This section outlines the implementation of text semantics-based product similarity using Word2Vec models. The goal is to recommend similar products based on their text descriptions.

* Average Word2Vec Product Similarity
* In this part, Word2Vec is employed to calculate the average vector representation of product titles. For each product title in the dataset, the average vector is computed using the pre-trained Word2Vec model. This representation allows for measuring semantic similarity between product titles.
* IDF Weighted Word2Vec for Product Similarity
* It focuses on enhancing the previous approach by incorporating Inverse Document Frequency (IDF) weighted Word2Vec. This technique assigns weights to words based on their importance in the entire corpus. The IDF-weighted vector representations of product titles are calculated, and product recommendations are made using these vectors.
* Weighted Similarity using Brand and Color
* In this section, it introduces additional features like brand, product type, and color to the similarity calculation. These categorical features are converted into numerical representations using Count Vectorization. The resulting vectors are concatenated with the previously calculated Word2Vec vectors. This composite vector is then used to recommend products, considering both textual and categorical attributes.

1. **Image-based Product Recommendation**

* Our recommendation engine incorporates image-based product recommendation using pre-trained Convolutional Neural Networks (CNNs). This approach leverages the visual features of apparel items to suggest visually similar products to users. The following steps detail the implementation of this approach:
* Utilization of Pre-trained Models
  + - We utilize pre-trained CNN models (such as VGG-16) to extract high-level image features. These models are pre-trained on large image datasets, enabling them to capture meaningful visual representations.
* Feature Extraction
  + - The images of apparel products are fed into the pre-trained CNN model to obtain a dense vector of features for each image. This vector encapsulates the image's visual characteristics in a high-dimensional space.
* Image Similarity Calculation
  + - Image similarity is measured using techniques like Euclidean or cosine distance. These measures quantify the similarity between the image feature vectors.
* Implementation and Code
  + - We utilize libraries such as Keras and TensorFlow for efficient image processing and feature extraction.
    - It also demonstrates the calculation of image similarity using different weights for various features, such as Word2Vec-based text features, brand, color, and image itself.
* 5.Visualizing Recommendations
  + - For a given input product, we calculate image-based similarities between that product and other products in the dataset.
    - The top-n visually similar products are selected based on the calculated image similarity scores.
    - Similarity metrics, such as Euclidean distance, quantify the similarity between the input product and recommended products.

Incorporating image-based recommendation enhances the diversity and accuracy of product suggestions, allowing users to discover visually appealing items that align with their preferences.

**Conclusion**

* when we are giving more weight to the BRAND we are getting maximum recommended product similar to query product brand same thing as when we are giving more weight to the image we are getting maximum recommended product similar to query product image
* When we have tried with different weight of image we are not slightly different not much different product with change in weight of the image

1. A/B Testing:

Further work, Conducting A/B Testing and how this project will be generalized for small business is still left.

# **EXPECTED OUTCOMES**

* Improved User Engagement:

We anticipate that the image-based recommendation approach will lead to increased user engagement, as visually similar products tend to capture users' attention more effectively. Higher engagement can manifest as increased CTR and longer browsing sessions.

* Enhanced Conversion Rates

By suggesting products that are visually appealing and closely aligned with users' preferences, we expect an improvement in conversion rates. Users are more likely to make purchases if they are presented with products that match their preferences.

* Diverse and Relevant Suggestions

The image-based approach should provide a wider variety of product recommendations, capturing nuanced visual features that might not be captured by text-based methods alone. This diversity can lead to improved user satisfaction and discovery.

* Small Business Benefit

The project's primary goal is to provide small businesses with an effective recommendation engine. We anticipate that our solution will help small businesses increase sales and customer satisfaction by offering personalized and visually appealing product suggstions.

# **REFRENCES**

1. Isinkaye, Folasade Olubusola, Y. O. Folajimi, and Bolande Adefowoke

Ojokoh. "Recommendation systems: Principles, methods, and

evaluation." Egyptian Informatics Journal 16.3 (2015): 261-273.

[CrossRef]

1. Röllecke, Felix Johannes, Arnd Huchzermeier, and David Schröder.

"Returning customers: The hidden strategic opportunity of returns

management." California Management Review 60.2 (2018): 176-203.

[CrossRef]