SAMSUNG

Big Data Course

Capstone Project
Final Report

I-Powered Product Recommendation System for E-commerce

OU _ BD2 _ GROUP 3

Hồ Ngọc Nhung

Bùi Ngọc Phương Linh

Phạm Thị Khánh Ly

Bùi Dạ Lý

Yoon Min

Phan Lê Nguyên

Trần Nhật Minh

14/08/2024

Contents

1. Introduction

1.1. Background Information

This project is centered around the development and implementation of a product recommendation system. The system uses machine learning techniques to predict and suggest items that users may be interested in. Due to challenges in sourcing official datasets, this project utilizes publicly available data from GitHub[1], focusing on the practical application of machine learning models with real-world data.

1.2. Motivation and Objective

The primary motivation behind this project is to develop a deeper understanding of product recommendation systems, which are essential tools in modern e-commerce. The objective is to implement, analyze, and compare various recommendation algorithms, including collaborative filtering, content-based methods, hybrid approaches, k-Nearest Neighbors (kNN), Random Forest, Support Vector Machines (SVM), and Logistic Regression. The goal is to uncover insights on optimizing these systems for superior performance and enhanced user satisfaction.

1.3. Members and Role Assignments

Project Manager: Hồ Ngọc Nhung

Responsible for overseeing the entire project, managing the timeline, coordinating tasks, and ensuring that the project stays on track.
System Integration and Implementation Engineer: Pham Thị Khánh Ly

In charge of integrating and deploying the developed models, ensuring smooth system operation, and making necessary adjustments during implementation.
Evaluation and Testing Coordinators: Bùi Dạ Lý, Bùi Ngọc Phương Linh

Tasked with evaluating the performance of the models, conducted thorough testing, and ensuring accuracy before final deployment.
Data Collection and Preprocessing Lead: Yoon Min

Responsible for leading the data collection process from various sources, cleaning, and preprocessing the data to prepare it for model development.
Model Development and Training Specialists: Trần Nhật Minh, Phan Lê Nguyên

Focused on developing, training, and optimizing machine learning models, specifically KNN and Random Forest, to ensure high performance.

1.4. Schedule and Milestones

Days 1-3: Project Planning and Setup
Initial project planning, task allocation, and setting up the necessary tools and environment.

Days 4-6: Data Collection
Gathering data from GitHub and other relevant sources.

Days 7-10: Data Preprocessing and Feature Engineering
Cleaning the data, extracting features, and preparing the dataset for model development.

Days 11-13: Model Development
Developing and training the KNN and Random Forest models.

Days 14-16: Model Evaluation and System Integration
Evaluating model performance, integrating models into the system, and testing.

Days 17-20: Deployment and Final AdjustmentsDeploying the final system, making necessary adjustments, and documenting the project.

2. Project Execution

2.1. Simulated Scenario Description

The project simulates an online retail environment where customers interact with various products. The goal is to build a recommendation system that predicts products a user might be interested in based on their previous interactions. Scenarios include user interactions, product metadata, and the implementation of recommendation algorithms.

2.2. Datasets Selection and Description

The dataset, selected from a publicly available repository, includes attributes related to user interactions with products. Key attributes include User ID, Product ID, Rating, Review Count, Category, Brand, Name, Image URL, Description, and Tags.

2.3. Data Ingestion Pipeline

The data ingestion pipeline involves several stages:

Data Loading: The dataset is loaded into a Pandas DataFrame.
Column Selection and Renaming: Relevant columns are selected and renamed for clarity.
Handling Missing Values: Missing values are handled by filling with appropriate defaults.
Data Integrity Checks: Duplicate records are identified and removed.

2.4. Data Transformation Processing

Data Loading and Initial Preparation

The initial step involved loading the dataset from a tab-separated file (walmart.tsv). The dataset was refined to include only relevant columns such as product ratings, reviews, categories, brands, descriptions, and tags. Missing values in critical fields were addressed by substituting them with default values or empty strings. Any duplicate entries were identified and removed to ensure data integrity.

Text Data Preprocessing

Textual data in columns like 'Product Category', 'Product Brand', 'Product Description', and 'Product Tags' underwent preprocessing to enhance its usability for analysis. This included converting text to lowercase, tokenizing, and removing stop words using natural language processing tools. The processed text from these columns was then combined into a single feature, known as 'Feature', which consolidates all relevant textual information for further analysis.

Feature Extraction with TF-IDF

To transform the textual 'Feature' column into a numerical format suitable for machine learning, the TF-IDF (Term Frequency-Inverse Document Frequency) method was applied. This technique quantifies the importance of each term in the documents, considering both term frequency and document frequency. The text data was transformed into a sparse matrix representation using this method, which facilitates subsequent clustering and modeling tasks.

Dimensionality Reduction and Clustering

Dimensionality reduction was performed to simplify the dataset and reduce computational complexity. Truncated Singular Value Decomposition (SVD) was used to reduce the dimensionality of the TF-IDF matrix. Following this, clustering was conducted using the KMeans algorithm. The optimal number of clusters was determined through the Elbow method, which evaluates how the variance within clusters changes as the number of clusters increases. Principal Component Analysis (PCA) was also utilized to determine the number of components that capture a significant portion of the variance in the data.

Data Transformation for Modeling
The preprocessed and transformed data was prepared for various machine learning models. Features selected for modeling included product categories, brands, descriptions, and tags. The dataset was split into training and test sets, and multiple models such as K-Nearest Neighbors, Random Forest, Support Vector Machine, Logistic Regression, Decision Trees, and Neural Networks were trained and evaluated for their performance.

2.5. Data Query and Insight

Model Training and Evaluation
After the data transformation and feature extraction processes, various machine learning models were trained to analyze and categorize the products. The models evaluated included:

K-Means Clustering: This model was used to group products into clusters based on their features. The number of clusters was determined using the Elbow method, which helps in identifying the optimal number of clusters by plotting the total within-cluster variance against different values of k. The clusters provide insight into the natural groupings within the product data.

Random Forest: A powerful learning ensemble model was used for classification tasks. It builds multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. This model was evaluated based on accuracy and was found to perform exceptionally well in categorizing products.

K-Nearest Neighbors (KNN): This algorithm classifies products based on the majority class among the nearest neighbors in the feature space. It was tested with different values of k to find the best-performing configuration.

Support Vector Machines (SVM): This model was used to find the optimal hyperplane that separates the different categories of products. The SVM with a linear kernel was chosen for its simplicity and effectiveness in classification tasks.

Logistic Regression: A statistical model that estimates probabilities using a logistic function. It was evaluated for its performance in categorizing products based on their features.

Decision Trees: This model builds a tree-like structure to make decisions based on feature values. It was used to understand how different features influence product categorization.
Model Selection for Recommendation System
Among the tested models, the Random Forest model demonstrates the highest accuracy in predicting product categories. Due to its robustness and superior performance, it was selected as the final model for the recommendation system.

Recommendation System Implementation
Content-Based Filtering:

Products were recommended based on their textual feature similarity. Using the cosine similarity measure, products similar to a given item were identified and recommended to users.
Collaborative Filtering:

A collaborative filtering approach was employed to suggest products based on user-item interactions. By analyzing user preferences and similarities, the system recommends items that similar users liked but the target user had not yet interacted with.
Hybrid Recommendation System:

The final recommendation system combines the results from both content-based and collaborative filtering approaches. This hybrid approach leverages the strengths of both methods to provide more accurate and diverse product recommendations.
Overall, the Random Forest model's effectiveness in product categorization and the integration of multiple recommendation techniques results in a robust and reliable recommendation system, enhancing user experience by delivering personalized and relevant product suggestions.

3. Results

3.1. Data Ingestion Scripts and Code

```
import pandas as pd


# Load data from TSV file
train_data = pd.read_csv('walmart.tsv', sep='\t')


# Select relevant columns
train_data = train_data[['Uniq Id', 'Product Id', 'Product Rating', 'Product Reviews
Count', 'Product Category', 'Product Brand', 'Product Name', 'Product Image Url',
'Product Description', 'Product Tags']]


# Display initial rows of data
train_data.head(3)


# Check for missing values
train_data.isnull().sum()


# Fill missing values
train_data['Product Rating'].fillna(0, inplace=True)
train_data['Product Reviews Count'].fillna(0, inplace=True)
train_data['Product Category'].fillna('', inplace=True)
train_data['Product Brand'].fillna('', inplace=True)
train_data['Product Description'].fillna('', inplace=True)


# Check for missing values after imputation
train_data.isnull().sum()


# Check for duplicate rows
train_data.duplicated().sum()
```

3.2. Data Transformation Scripts and Code

```
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import TruncatedSVD


# Load SpaCy model
nlp = spacy.load("en_core_web_sm")


def clean_and_extract_tags(text):
    doc = nlp(text.lower())
    tags = [token.text for token in doc if token.text.isalnum() and token.text not in
STOP_WORDS]
```

```
    return ', '.join(tags)

# Process text data
columns_to_extract_tags_from = ['Product Category', 'Product Brand', 'Product
Description', 'Product Tags']
for column in columns_to_extract_tags_from:
    train_data[column] = train_data[column].apply(clean_and_extract_tags)

# Combine text columns into a single feature
train_data['Feature'] = train_data[columns_to_extract_tags_from].apply(lambda row:
','.join(row), axis=1)

# TF-IDF Vectorization
from nltk.stem.snowball import SnowballStemmer
import nltk
import re

nltk.download('punkt')
nltk.download('stopwords')

stemmer = SnowballStemmer("english")

def tokenize_and_stem(text):
    tokens = [word for word in nltk.word_tokenize(text)]
    fillter_tokens = [token for token in tokens if re.search("[a-zA-Z]", token)]
    stems = [stemmer.stem(word) for word in fillter_tokens]
    return stems

tfidf_vectorizer = TfidfVectorizer(max_df=0.7, min_df=0.1, ngram_range=(1, 3),
stop_words="english", use_idf=True, tokenizer=tokenize_and_stem)
tfidf_matrix = tfidf_vectorizer.fit_transform([x for x in train_data['Feature']])

# Dimensionality Reduction
svd = TruncatedSVD(n_components=51, random_state=42)
reduced_data = svd.fit_transform(tfidf_matrix)
```

3.3. Description and Sample of Transformed Datasets

Description

The transformed dataset has been enriched with features that are derived from the text columns of the original dataset. specifically, these features include:

Uniq Id: A unique identifier for each product.

Product Id: The unique identifier for the product.

Product Rating: The rating given to the product, with missing values filled in as 0.

Product Reviews Count: The number of reviews for the product, with missing values filled in as 0.

Product Category: The category of the product, with non-alphanumeric characters removed and text cleaned.

Product Brand: The brand of the product, with non-alphanumeric characters removed and text cleaned.
Product Name: The name of the product.
Product Image Url: The URL for the product image.
Product Description: A description of the product, with non-alphanumeric characters removed and text cleaned.
Product Tags: Tags associated with the product, with non-alphanumeric characters removed and text cleaned.
Feature: A combined feature of cleaned Product Category, Product Brand, Product Description, and Product Tags used for further processing.
categ_product: The category of the product, assigned via clustering.

Sample of Transformed Datasets

	Uniq Id	Product Id	Product Rating	Product Reviews Count	Product Category	Product Brand	Product Name	Product Image Url	Product Description	Product Tags	Feature	categ_product
0	1705736792d82aa2f2d3caf1c07c53f4	2e17bf4acecdece67fc00f07ad62c910	0.0	0.0	premium, beauty, premium, makeup, premium, nai...	opi	OPI Infinite Shine, Nail Lacquer Nail Polish, ...	https://i5.walmartimages.com/asr/0e1f4c51-c1a4...		opi, infinite, shine, nail, lacquer, nail, pol...	premium, beauty, premium, makeup, premium, nai...	3
1	95a9fe6f4810fcfc7ff244fd06784f11	076e5854a62dd283c253d6bae415af1f	0.0	0.0	beauty, hair, care, hair, color, auburn, hair,...	easy	Nice n Easy Permanent Color, 111 Natural Mediu...	https://i5.walmartimages.com/asr/9c8e42e4-13a5...	pack, 3, pack, 3, upc, 381519000201, beautiful...	nice, n, easy, permanent, color, 111, natural,...	beauty, hair, care, hair, color, auburn, hair,...	4
2	8d4d0330178d3ed181b15a4102b287f2	8a4fe5d9c7a6ed26cc44d785a454b124	4.5	29221.0	beauty, hair, care, hair, color, permanent, ha...	clairol	Clairol Nice N Easy Permanent Color 7/106A Nat...	https://i5.walmartimages.com/asr/e3a601c2-6a2b...	clairol, nice, n, easy, permanent, color, give...	clairol, nice, n, easy, permanent, color, natu...	beauty, hair, care, hair, color, permanent, ha...	2

Feature: Represents a concatenation of cleaned and processed text data from various columns, making it suitable for feature extraction techniques like TF-IDF.
categ_product: Indicates the cluster or category assigned to each product based on the K-Means clustering algorithm.

3.4. Data Visualization of Query Results

The dataset was analyzed to determine the number of unique users, items, and ratings. This provides an understanding of the dataset's scope and helps in evaluating the diversity of the product and user base.

Number of Unique Users: [num_users]
Number of Unique Items: [num_items]
Number of Unique Ratings: [num_ratings]

3.4.1 Heatmap of User Ratings

A heatmap was generated to visualize the distribution of user ratings across different products. This heatmap illustrates the frequency of ratings provided by users and helps identify patterns or anomalies in user behavior.

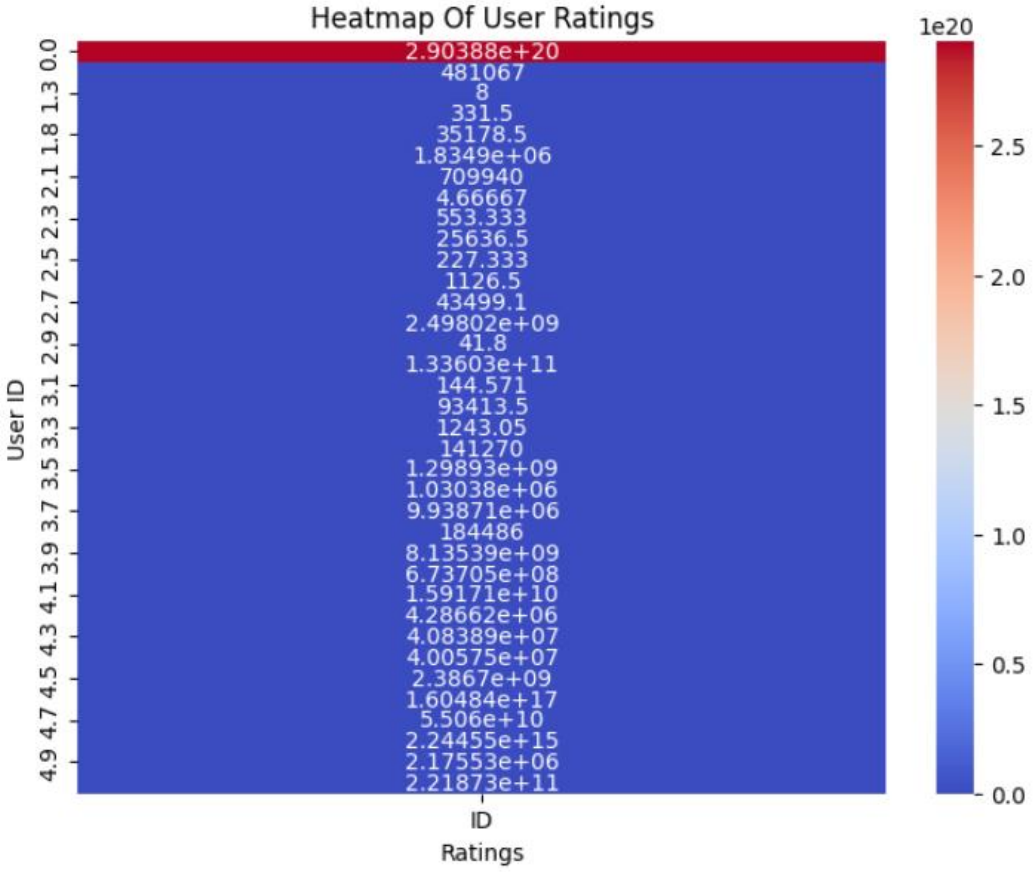

Figure 1: Heatmap of User Ratings

Figure 1: Heatmap showing the distribution of user ratings. Each cell represents the number of users who have given a specific rating to a product.

3.4.2 Distribution of Interactions

To understand the interaction patterns, histograms were plotted showing the distribution of interactions per user and per item.

Distribution of Interactions Per User: This histogram displays how frequently users interact with items. It highlights the variation in user engagement levels.

Distribution of Interactions Per Item: This histogram illustrates the number of interactions each item receives, revealing the popularity of items within the dataset.

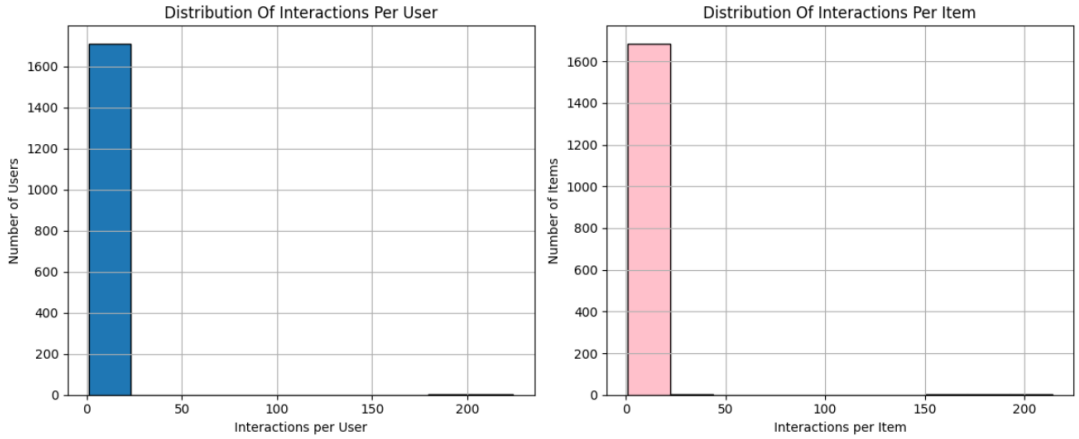

Figure 2: Distribution of Interactions Per User and Item

Figure 2a: Histogram showing the distribution of interactions per user.
Figure 2b: Histogram showing the distribution of interactions per item.

3.4.3 Most Popular Items

A bar chart was created to display the top 10 most popular items based on interaction counts. This visualization helps in identifying which products are most favored by users.

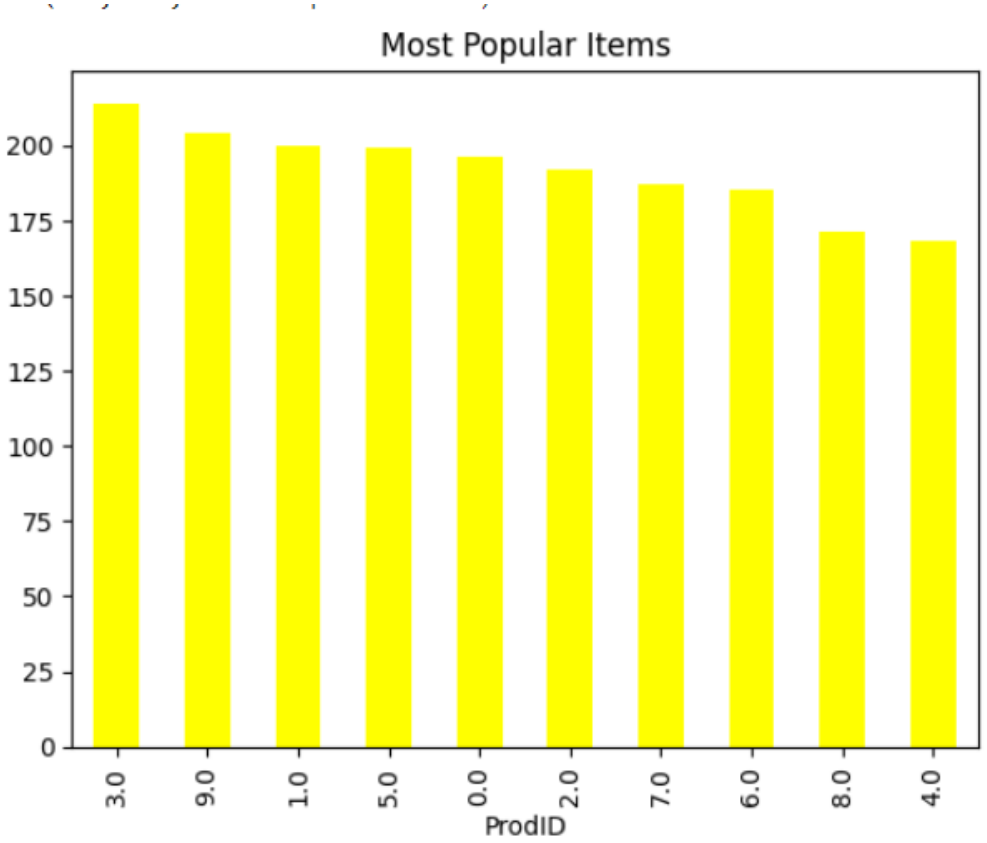

Figure 3: Most Popular Items

Figure 3: Bar chart representing the top 10 most popular items.

3.4.4 Most Rated Items

Another bar chart shows the distribution of ratings for different products. This visualization highlights which items have been rated the most, providing insight into user engagement with particular products.

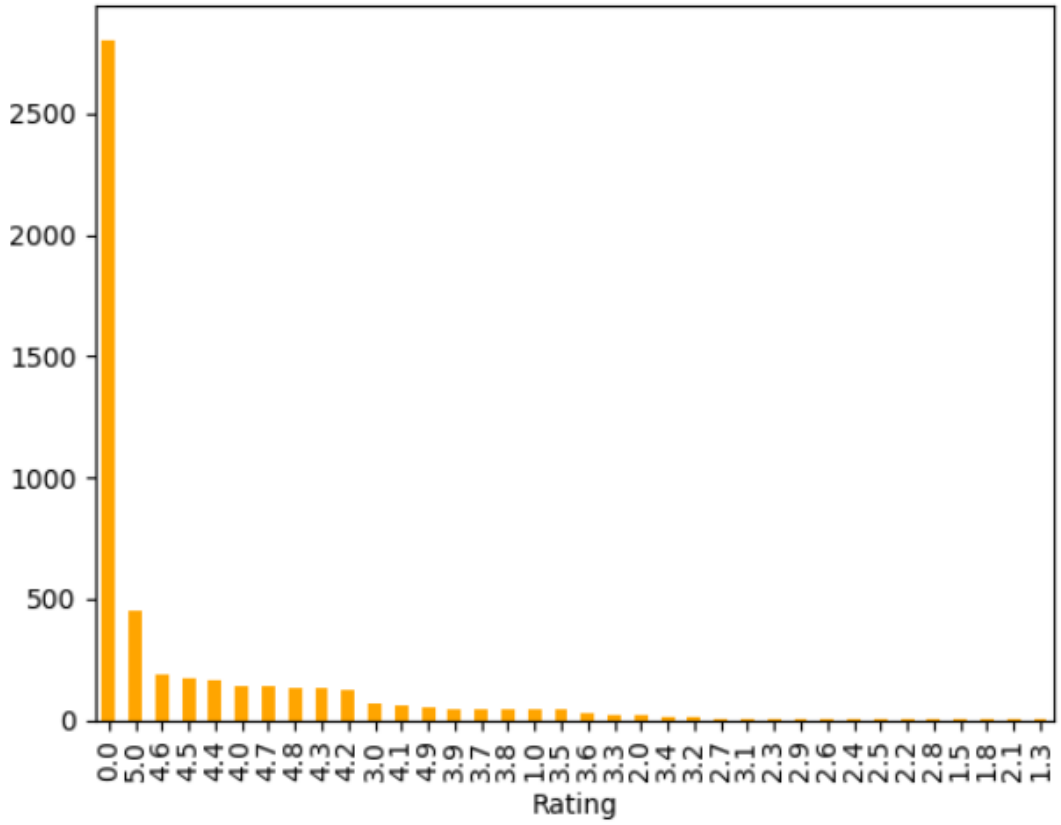

Figure 4: Most Rated Items

Figure 4: Bar chart depicting the distribution of ratings for various items.

4. Projected Impact

4.1. Accomplishments and Benefits

Accomplishments:

Comprehensive Data Processing:

Successfully cleaned and preprocessed the dataset by handling missing values, eliminating duplicates, and renaming columns for clarity.
Extracted and normalized textual data for improved analysis and feature extraction.
Effective Data Visualization:

Generated insightful visualizations that highlight user interactions, product popularity, and rating distributions.
Created heatmaps and histograms to provide a clear understanding of user engagement and product performance.
Model Development and Evaluation:

Implemented and evaluated several machine learning models, including K-Means Clustering, Random Forest, K-Nearest Neighbors, Support Vector Machines, Logistic Regression, and Decision Trees.
Selected the Random Forest model for its superior accuracy in categorizing products and generating recommendations.
Recommendation System Implementation:

Developed a robust recommendation system utilizing both content-based and collaborative filtering approaches.
Integrated a hybrid recommendation model to leverage the strengths of both methods, enhancing the quality of product suggestions.
Benefits:

Enhanced User Experience:

The recommendation system provides personalized and relevant product suggestions improving, user satisfaction and engagement.
Visualizations offer valuable insights into user behavior and product performance, aiding in data-driven decision-making.
Improved Product Categorization:

The Random Forest model's accuracy in categorizing products ensures that users receive accurate and meaningful recommendations.
Informed Business Decisions:

Data visualizations and insights support strategic decisions regarding product placement, marketing strategies, and inventory management.

4.2. Future Improvements

1. Model Refinement:

Experiment with Advanced Models: Explore and integrate more advanced machine learning models, such as neural networks or gradient boosting algorithms, to enhance the accuracy and performance of the recommendation system.
Hyperparameter Tuning: Perform extensive hyperparameter tuning to optimize model performance and improve recommendation quality.
2. Enhanced Feature Engineering:

Incorporate Additional Features: Integrate additional features such as user demographics, product metadata, and seasonal trends to provide more contextually relevant recommendations.
Textual Data Enrichment: Use advanced natural language processing techniques to extract more meaningful insights from product descriptions and reviews.
3. Scalability and Performance:

Optimize Computational Efficiency: Implement optimization techniques to handle large-scale datasets and reduce computational costs.
Real-Time Recommendations: Develop real-time recommendation capabilities to provide users with immediate and dynamic suggestions based on their latest interactions.
4. User Feedback Integration:

Incorporate User Feedback: Implement mechanisms to gather user feedback on recommendations and use this data to continuously improve the recommendation algorithms.
A/B Testing: Conduct A/B testing to evaluate the effectiveness of different recommendation strategies and refine the system based on empirical results.
5. Expand System Capabilities:

Multi-Channel Integration: Extend the recommendation system to work across multiple channels (e.g., web, mobile) for a seamless user experience.
Cross-Selling and Upselling: Develop strategies for cross-selling and upselling based on user preferences and purchase history.

5. Team Member Review and Comment

NAME	REVIEW and COMMENT
Hồ Ngọc Nhung	This project marked my first experience as a Project Manager, and it was a significant challenge. As a new endeavor for all of us, the lack of prior experience added to the complexity. Despite these hurdles, we successfully completed the project on time. I am sincerely pleased with the outcome and deeply grateful for the hard work and dedication of each team member.
Bùi Ngọc Phương Linh	It was my first encounter with Big Data, beginning with only a limited understanding and experience. However, through the course and the final project, I gained valuable insights and practical skills. I take pride in our accomplishments and am deeply grateful for the guidance and collaborative effort that made them possible. This experience marks a significant milestone in my educational journey.
Phạm Thị Khánh Ly	Working on a Big Data project has been an incredible learning experience. I've had the opportunity to dive deep into the world of data analysis and explore various tools and techniques. While there were challenges along the way, such as dealing with large datasets and complex algorithms, I found the process incredibly rewarding. Discovering hidden insights and patterns within the data was truly fascinating, and I'm excited to apply these skills to future projects.
Bùi Dạ Lý	This was my first time learning about Big Data, I started with limited knowledge and experience. However, the course and the final project gave me valuable insights and practical skills. I'm proud of our achievements and thankful for the guidance and teamwork that made it possible. This experience has been a significant milestone in my learning journey.
Yoon Min	This project not only broadens my knowledge of data science but also enhances my teamwork skills. It helps me apply new knowledge in practice and improve my ability to coordinate effectively with team members, which is very beneficial for my career.

Phan Lê Nguyên	This is the first Big Data project I have participated in. Although my knowledge and experience are still limited, I have learned a lot of new knowledge that will be helpful for my future journey.
Trần Nhật Minh	It was a journey of growth, where every challenge became an opportunity to learn and innovate. This experience has strengthened my skills and confidence, and I look forward to applying these lessons to future endeavors.

APPENDIX: Data Source

[1] N. Saeed, "E-Commerce Recommendation System - Machine Learning Product Recommendation System," GitHub Repository, 2023. [Online]. Available: https://github.com/611noorsaeed/E-Commerece-Recommendation-System-Machine-Learning-Product-Recommendation-system-/tree/main. [Accessed: Aug. 14, 2024].

6. Instructor Review and Comment

CATEGORY	SCORE	REVIEW and COMMENT
IDEA	__/10	
APPLICATION	__/30	
RESULT	__/30	
PROJECT MANAGEMENT	__/10	
PRESENTATION & REPORT	__/20	

TOTAL	__/100	