FAST NATIONAL UNIVERSITY

Advanced Big Data Analytics DS5001 Semester Project

Product Recommendation System Using Alternating Least Squares (ALS)

Group Members:

Moosa Raza
24k-8071

M. Baasil
24k-8070

Abstract

With the rapid expansion of e-commerce platforms, users are exposed to an overwhelming number of products, making personalized recommendations essential. This project presents a scalable product recommendation system using the Alternating Least Squares (ALS) algorithm implemented on Apache Spark. The Instacart Market Basket Analysis dataset is used to model user to product interactions based on purchase history. The system is built using Databricks Community Edition and Spark MLlib to demonstrate big data processing and collaborative filtering at scale. Model performance is evaluated using Root Mean Square Error (RMSE), achieving a value of approximately 4.2, indicating effective prediction of user preferences.

1. Introduction

Recommendation systems play a vital role in modern digital platforms by helping users discover relevant products and content. Traditional recommendation approaches struggle to scale when dealing with massive, sparse datasets. Big data technologies such as Apache Spark enable distributed processing and efficient model training on large-scale datasets.

This project focuses on building a collaborative filtering recommendation system using the ALS algorithm, which is widely adopted in industry for scalable recommender systems. The goal is to demonstrate how big data tools can be applied to solve a real-world personalization problem.

2. Problem Statement

Online retail platforms face the challenge of information overload, where users must choose from thousands of available products. Without personalization, user experience and engagement suffer. The problem addressed in this project is to design a scalable system that can analyze large volumes of transaction data and recommend products that users are likely to purchase in the future.

3. Dataset Description

The project uses the **Instacart Market Basket Analysis dataset**, a publicly available dataset from Kaggle. The dataset contains millions of grocery purchase transactions, including user identifiers, product identifiers, and order information.

- **Dataset size:** Over 30 million records
- **Data characteristics:** Large-scale, sparse, and transactional
- **Type of data:** Implicit feedback (purchase history)

The dataset qualifies as big data due to its size and complexity, making distributed processing necessary.

4. Tools and Technologies

The following tools and technologies were used:

- **Apache Spark:** Distributed data processing framework
- **Spark MLlib:** Machine learning library used for ALS
- **Databricks Community Edition:** Cloud-based big data platform
- **Python (PySpark):** Programming language for implementation
- **Spark SQL:** Querying and transformation of data

5. Methodology

The project follows a structured big data analytics pipeline:

1. **Data Ingestion:**
   The dataset was accessed as distributed Spark tables in Databricks.
2. **Data Preprocessing:**
   User–product interactions were constructed by aggregating purchase frequency. Missing and irrelevant values were filtered, and identifiers were indexed for ALS compatibility.
3. **Model Training:**
   A collaborative filtering model was trained using the ALS algorithm with implicit feedback. The dataset was split into training (80%) and testing (20%) sets.
4. **Evaluation:**
   Model performance was evaluated using Root Mean Square Error (RMSE).

5. **Recommendation Logic:**
   Product recommendations were derived by ranking predicted interaction scores generated by the ALS model.

6. Model Implementation

The ALS model was configured with the following parameters:

- Implicit feedback enabled
- Rank = 10
- Regularization parameter = 0.1
- Maximum iterations = 10

Purchase frequency was used as interaction strength. The model learned latent factors representing user preferences and product characteristics.

7. Evaluation and results

Model performance was evaluated using RMSE on the test dataset.

- **RMSE ≈ 4.20**

This value indicates that the model can reasonably predict user–product interaction strength in an implicit feedback setting. Sample prediction outputs showed higher scores for products with stronger predicted user preference, which were used as recommendations.

9. Conclusion

This project successfully demonstrates the application of big data technologies to build a scalable recommendation system. Using Apache Spark and ALS, a collaborative filtering model was trained on a large real-world dataset and evaluated effectively. The results show that big data frameworks are well-suited for solving personalization problems in modern e-commerce environments.