Amazon Data Mining Project

1. K-Means Clustering (Customer Segmentation)

What is it?
K-Means is an unsupervised machine learning algorithm used to group similar customers together based o

How it Works in Your Project:
- You select relevant features like discounted_price, actual_price, rating, and rating_count.
- Standardization: The data is standardized using StandardScaler() to make sure all features contribute equ
- The K-Means algorithm finds 3 clusters (customer groups) by minimizing intra-cluster distances.
- The final output is a scatter plot showing clusters based on discounted price vs. actual price.

Why is this Useful?
- Helps Amazon understand different customer purchasing patterns.
- Can be used for personalized recommendations and marketing strategies.

2. Exploratory Data Analysis (EDA)

What is it?
EDA helps in understanding data through visualizations and statistics before applying machine learning.

Steps in Your Project:
- Distribution Analysis: Histogram of discounted_price to see price distribution.
- Scatter Plots: Helps visualize relationships, e.g., actual_price vs. discounted_price.
- Correlation Matrix: Heatmap to analyze how variables are related (rating, price, etc.).

Why is this Useful?
- Identifies trends, anomalies, and patterns before applying machine learning models.
- Helps in feature selection and data cleaning.

3. User Behavior Analysis

What is it?
Analyzing how users interact with products, including:
- Reviews (review_content length distribution).
- Average Rating by Category (finding the most loved categories).
- Most Purchased Products (based on rating_count).

Why is this Useful?
- Helps in improving user experience.
- Can be used to optimize Amazon's recommendation system.

4. Association Rule Mining (Market Basket Analysis)

What is it?
Finding patterns in what customers frequently buy together using:
- Apriori Algorithm (used in your project).
- Generating association rules (e.g., "People who buy X also buy Y").

How it Works in Your Project:
- The apriori function finds frequent itemsets.
- association_rules() generates rules based on support, confidence, and lift.

Why is this Useful?
- Helps in cross-selling (Amazon recommends complementary products).
- Improves personalized recommendations.

Final Summary

Your project covers essential data mining techniques for Amazon:
- K-Means: Segments customers based on purchase behavior.
- EDA: Helps explore and understand data visually.
- User Behavior Analysis: Reveals buying patterns and preferences.
- Association Rule Mining: Finds frequently bought product combinations.

This information can help Amazon optimize pricing, marketing, and product recommendations!

Project by
Yashvardhan Kumar
Student at Birla Open Minds International School