Data Mining
Group 10: Qixi Huang, Jialong Li, Vedant Paithane, Darwin Ye

Project Proposal

Business Questions:

Product Recommendation Optimization

How can we improve the accuracy and relevance of product recommendations for customers based on their browsing and purchase history? This question addresses the development and refinement of recommender systems. By leveraging association rule mining and clustering analysis, you can identify patterns in customer behaviour and product associations. This would involve analyzing customer reviews, ratings, and purchasing patterns to understand preferences and suggest products that are likely to be of interest to similar customer segments.

Market Basket Analysis for Cross-Selling Opportunities

What are the potential product bundles that can be offered to customers to increase cross-selling opportunities? This question focuses on the application of association rule mining to discover products frequently purchased together. By analyzing the 'category', 'product_name', and 'review_content' fields, you can identify products that are often bought in combination, enabling targeted marketing strategies and the creation of bundled offers. This analysis can reveal insights into customer purchasing behaviour and preferences, facilitating more effective cross-selling and upselling strategies.

Motivation:

We are eager to apply our course learnings to a practical business challenge, emphasizing diverse data visualization techniques and advanced modeling. Opting for the Amazon sales

dataset enables us to grapple with a real-world scenario and implement recommender systems and association mining. Beyond the technical aspects, solving this problem holds significant real-world value. In the e-commerce landscape, effective recommender systems enhance user experience, drive sales, and foster customer loyalty. This project not only provides hands-on experience with intricate datasets but also addresses a problem statement with tangible implications for businesses, showcasing the practical relevance of our skills in data science and machine learning.

Methodology:

Dimensionality Reduction

Apply PCA, t-sne to visualize the dataset in lower-dimensional space. Identify the important features, check for the clusterability of the dataset, and avoid the curse of dimensionality.

Clustering Analysis

Implement clustering algorithms, such as K-means+, hierarchical clustering, Gaussian Mixture Model, and BIRCH to put similar products in the same cluster. And evaluate the result of clustering analysis properly.

Association Rule Mining

Use Association Rule Mining algorithms to uncover the relationships between features in the dataset. Identifying high-probability products, high-probability customers, and across effects of products.

NLP

NLP is not our focus of the project, but we don't exclude the possibility of using some NLP technique to process the "review_content" feature, such as cosine similarity.

Dataset

This project focuses on a dataset of over 1,000 Amazon product listings, featuring a blend of product details and customer reviews. Key features include product

identifiers, names, categories, pricing, ratings, and extensive review content. There are currently 16 features in this dataset, and we intend to conduct detailed text analysis of the 'about_product', 'category', and 'review_content' fields, aiming to significantly expand the feature space for richer insights. We believe this e-commerce dataset is an excellent fit for our projects, as it closely resembles a highly realistic use-case scenario in the realms of data mining and recommendation systems. By analyzing this dataset, we can gain insights that are directly applicable to real-world e-commerce challenges, including understanding customer preferences, improving product recommendations, and tailoring marketing strategies.