CUSTOMER PURCHASE BEHAVIOR PREDICTION & MARKET BASKET ANALYSIS

Farah Kareem (23K-8045)
Maham Tariq (23K-7804)
Fariha Hashmi (23k-8036)

Master of Data Science

Master of Data Science at the National University of Computer & Emerging Sciences

Department of Computer Science

National University of Computer & Emerging Sciences

2025

Introduction

In the fast-growing online grocery sector, understanding customer purchasing behavior is key to enhancing user experience, optimizing inventory, and increasing sales. This project uses Big Data technologies to analyze Instacart's extensive dataset, discover product associations, and predict future purchases.

Problem Statement

E-commerce platforms handle millions of transactions daily. The challenge is to process and analyze this large volume of data to:

a) Identify frequent purchase patterns
b) Recommend products users are likely to buy
c) Segment customers based on shopping behavior

Objectives

a) Analyze user behavior and purchase trends
b) Identify frequently bought-together items using FP-Growth
c) Predict next likely product using Collaborative Filtering (ALS)
d) Classify reorder probability using machine learning
e) Use Apache Spark for scalable data processing
f) Visualize insights to support business decision-making
g) Use PageRank on co-purchased product graphs

Dataset Overview

Source: Instacart Online Grocery Shopping Dataset (Kaggle)
Scale:

a) 3.4 million orders
b) 30 million product-order rows
c) 50K unique products
d) 200K+ users
e) 21 departments, 134 aisles

Files Used:

a) orders.csv, products.csv, departments.csv, aisles.csv
b) order_products__train.csv, order_products__prior.csv

Tools & Technologies

a) Apache Spark (PySpark): Distributed data processing
b) Google Colab : Prototyping and analysis
c) Spark MLlib: FP-Growth, ALS, Random Forest
d) Matplotlib / Seaborn: Visualization
e) NetworkX & Matplotlib: Graph-based analysis and visualization

f) PageRank Algorithm: Used for graph-based product ranking

Methodology

a) Data Preprocessing
- Load all CSVs into Spark DataFrames
- Merge product/order metadata with transactions

b) Feature Engineering
- Calculate user-specific and product-specific features like Order frequency, Reorder ratio, Add-to-cart sequence, and Time trends

c) Market Basket Analysis
- Applied FP-Growth to discover frequent itemsets and generate association rules. This method efficiently generates association rules from large transaction datasets without needing to generate candidate sets like Apriori

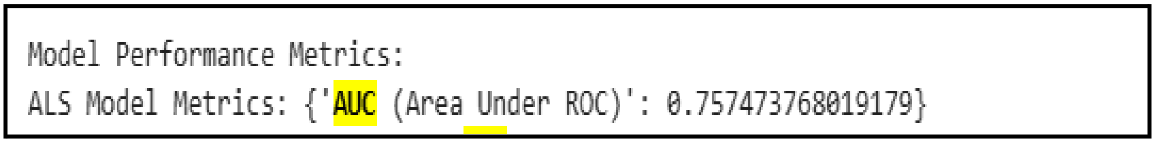
d) Predictive Modeling
A. Recommendation System
- ALS (Alternating Least Squares) to predict next product
B. Reorder Prediction
- Classification model using Random Forest

e) Visualization
- Plots and heatmaps for popular items, reorder trends, and associations

f) Graph-Based Recommendation using PageRank
- Constructed a co-purchase product graph where edges represent frequently co-bought products (count > 20,000)
- Nodes represent unique products
- Applied a normalized weighted PageRank algorithm to rank products based on their influence in the co-purchase network
- Used PySpark for scalable computation and NetworkX + Matplotlib for visualization

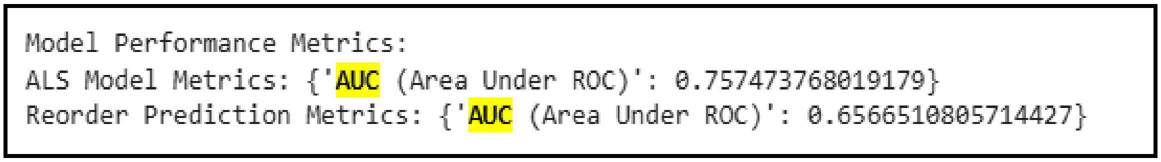
Model Performance & Evaluation

a) ALS Recommendation Model: Indicates strong capability in ranking relevant product recommendations.



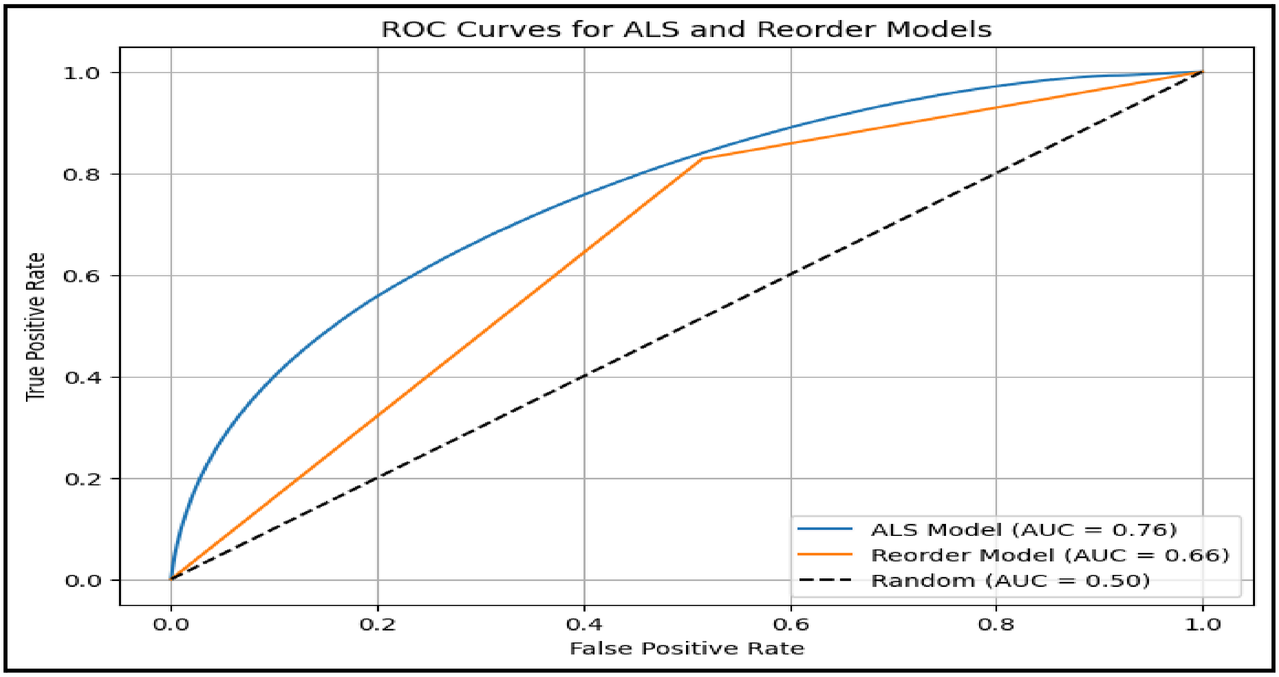
```
Model Performance Metrics:
ALS Model Metrics: {'AUC (Area Under ROC)': 0.757473768019179}
```

b) Reorder Classification Model : Shows decent performance in predicting whether a user will reorder a product.



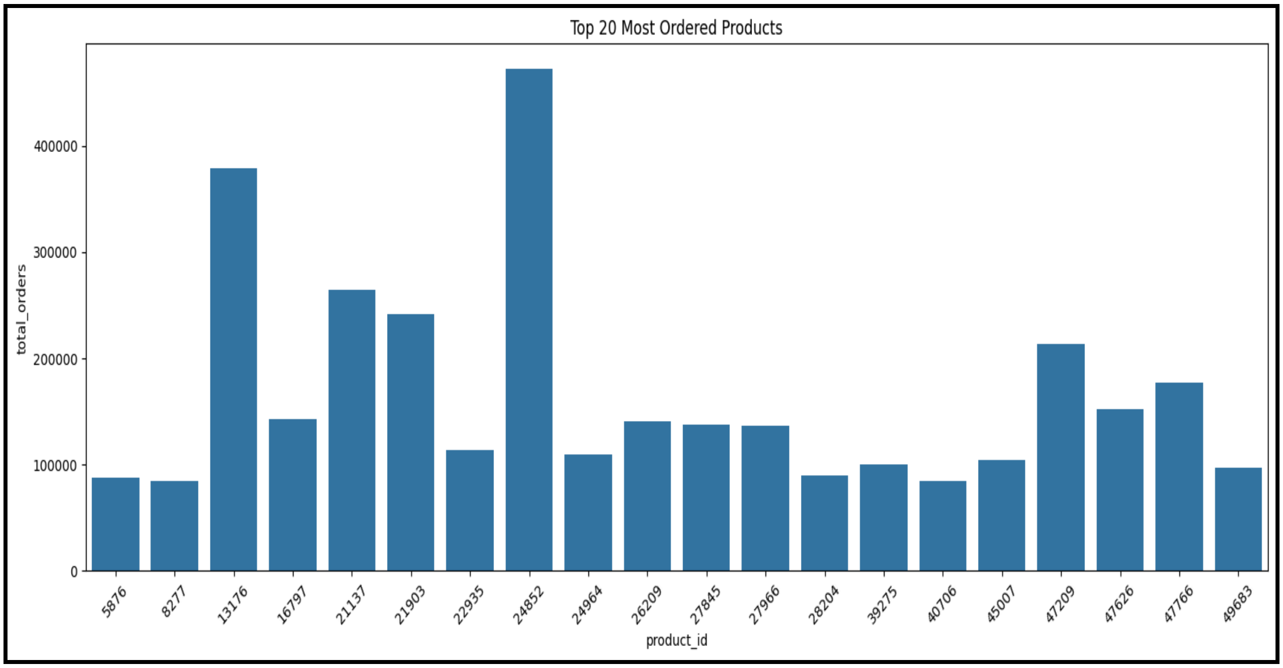
```
Model Performance Metrics:
ALS Model Metrics: {'AUC (Area Under ROC)': 0.757473768019179}
Reorder Prediction Metrics: {'AUC (Area Under ROC)': 0.6566510805714427}
```

c) ROC Curves : Below are the ROC curves for both models

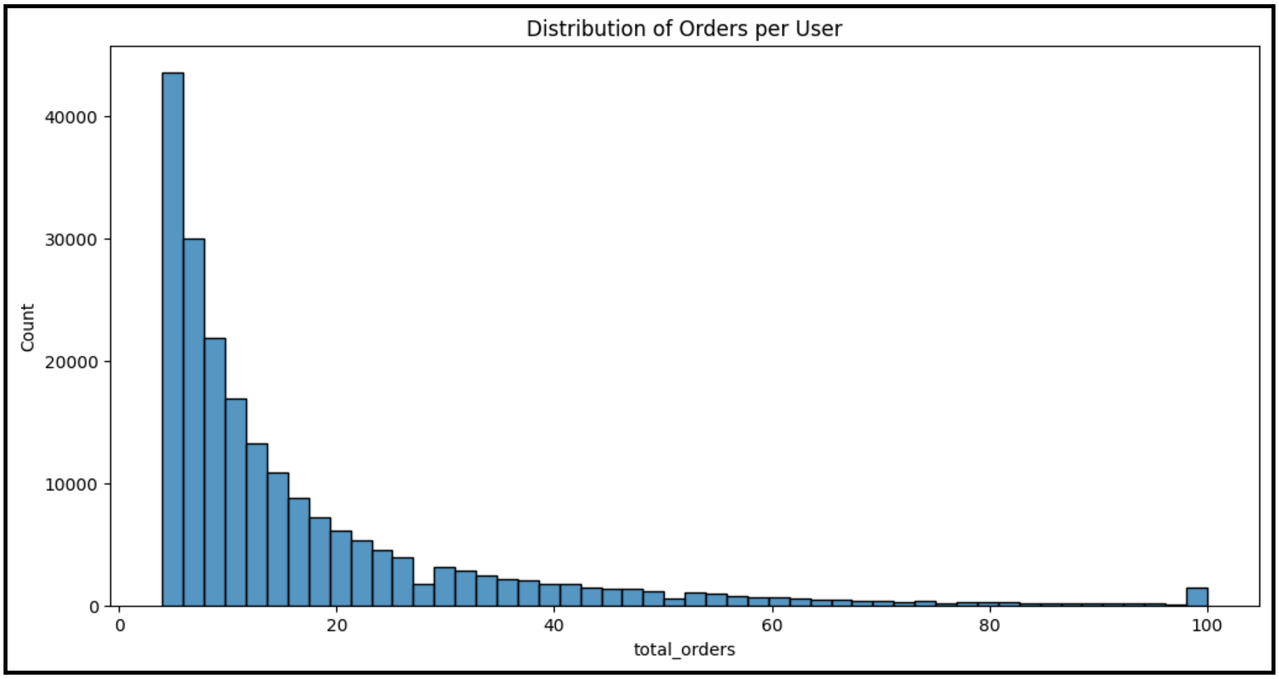


Visual Insights & Association Analysis

a) Top 20 Most Ordered Products: Visualizes product popularity across all orders.

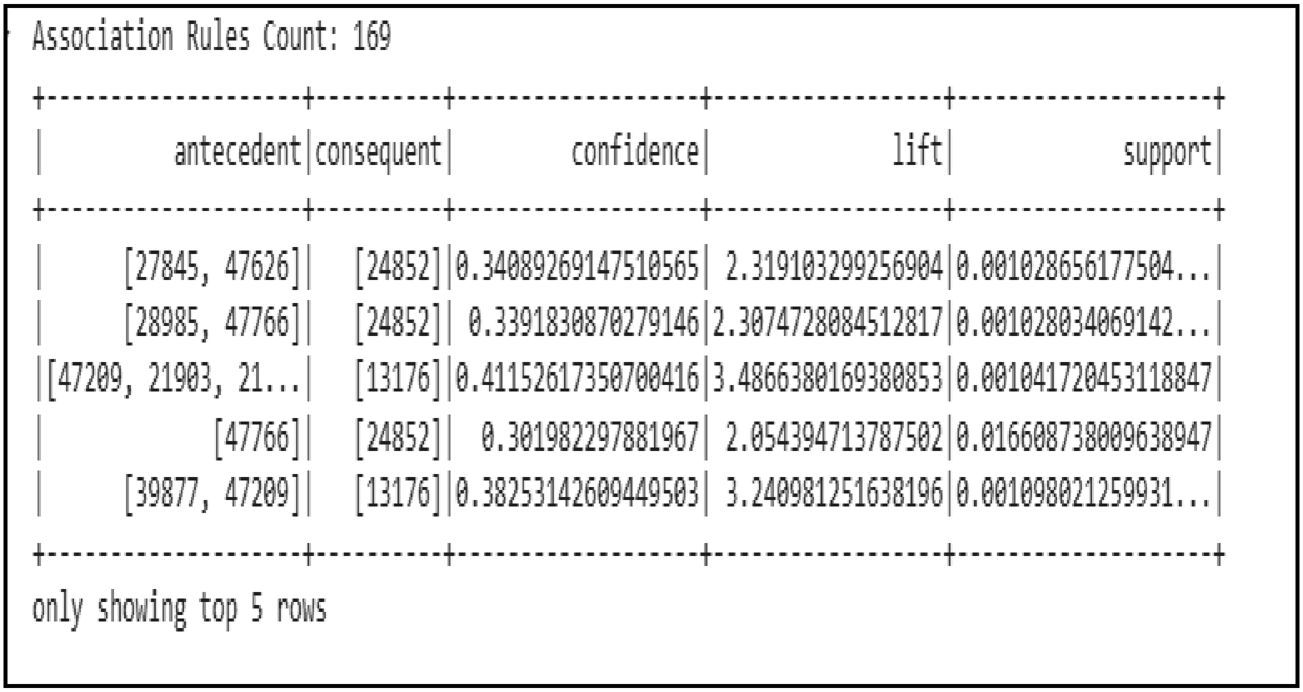


b) Distribution of Orders per Product: Highlights long-tail pattern in product ordering behavior.



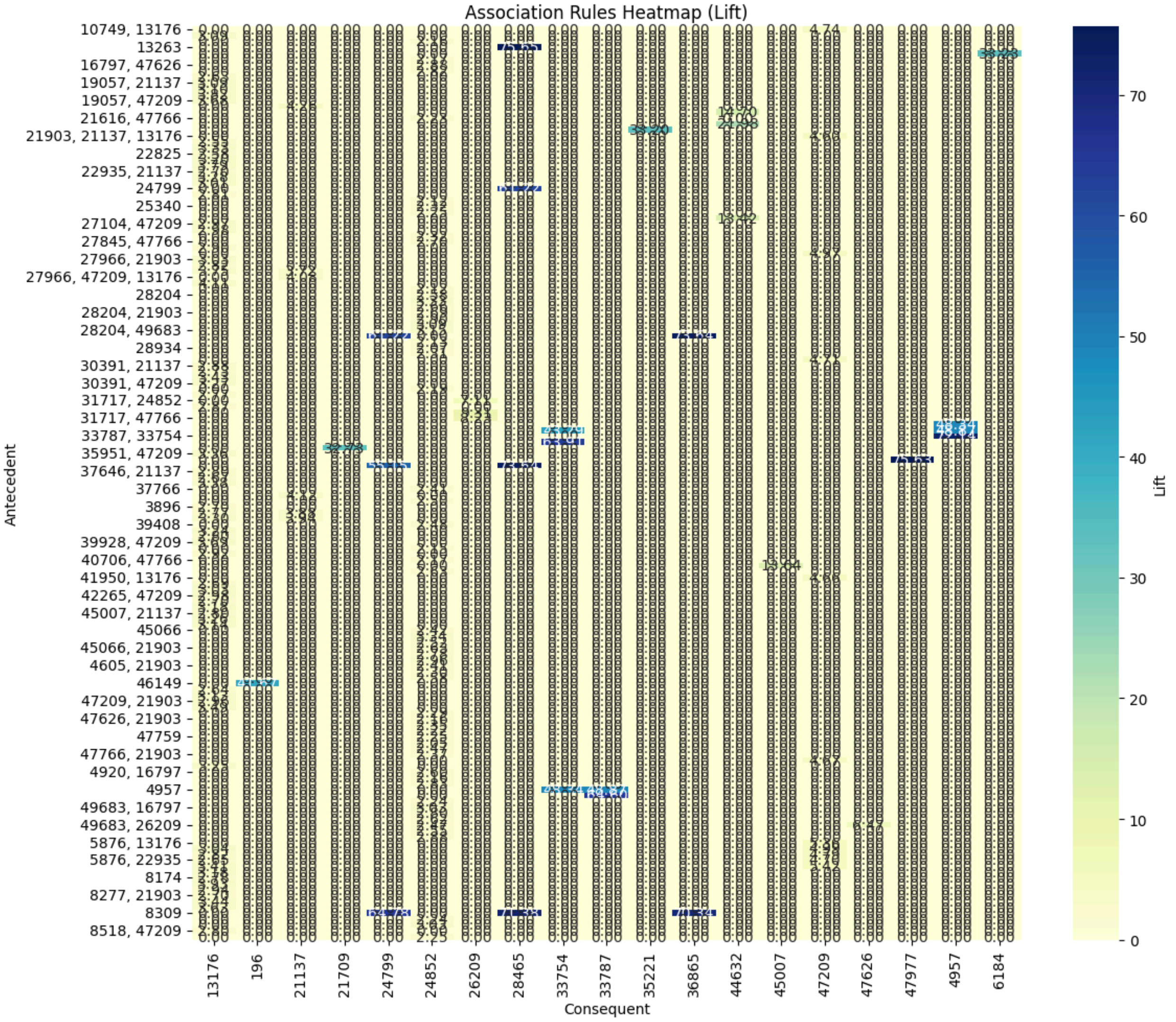
c) Frequent Itemsets (Market Basket Analysis):Found 169 association rules using FP-Growth

d) Sample Association Rules Table (Top 5)

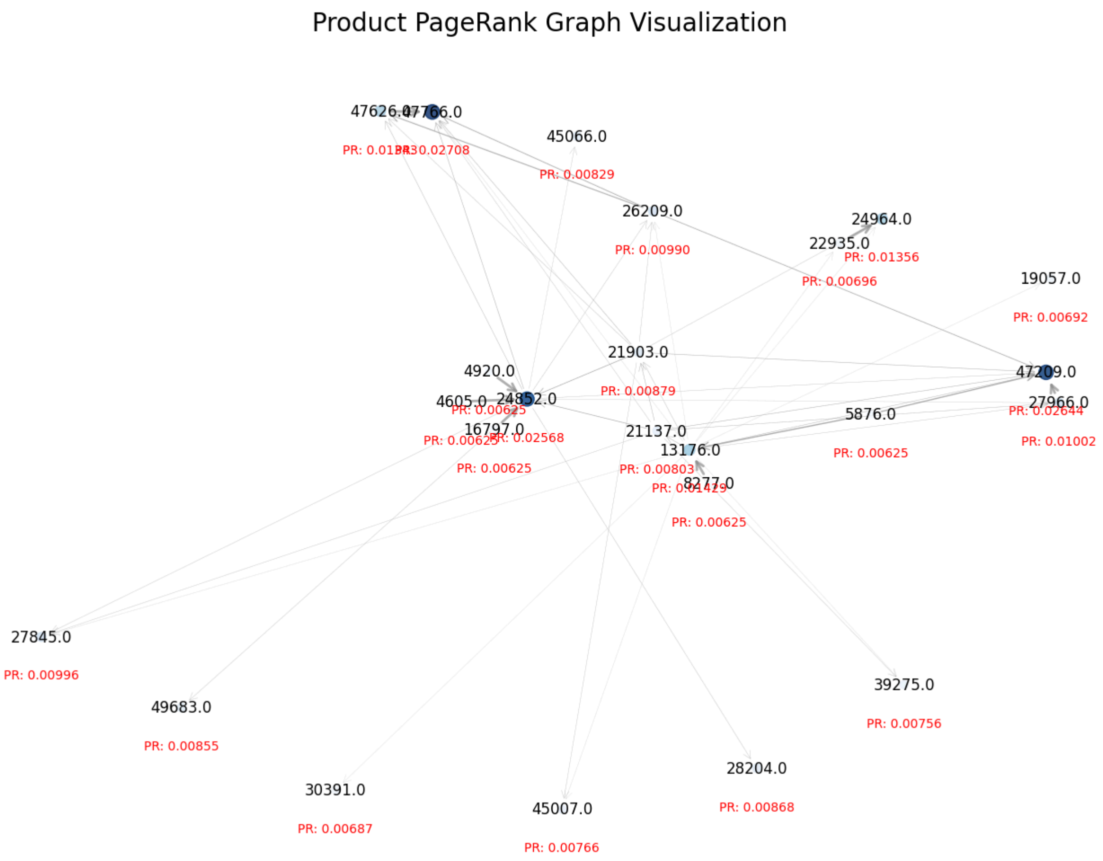


```
Association Rules Count: 169
+--------------------+----------+-------------------+------------------+--------------------+
|          antecedent|consequent|         confidence|              lift|             support|
+--------------------+----------+-------------------+------------------+--------------------+
|      [27845, 47626]|   [24852]|0.34089269147510565| 2.319103299256904|0.001028656177504...|
|      [28985, 47766]|   [24852]| 0.3391830870279146|2.3074728084512817|0.001028034069142...|
|[47209, 21903, 21...|   [13176]|0.41152617350700416|3.4866380169380853|0.001041720453118847|
|             [47766]|   [24852]|  0.301982297881967| 2.054394713787502|0.016608738009638947|
|      [39877, 47209]|   [13176]|0.38253142609449503| 3.240981251638196|0.001098021259931...|
+--------------------+----------+-------------------+------------------+--------------------+
only showing top 5 rows
```

e) Association Heatmap: Visualizes strength of product pair associations



f) Product PageRank Graph Visualization



Success Metrics

a) Precision / Recall: ≥ 75% for reorder classification

b) Lift in Recommendations: Improved relevance

c) Performance: Efficient scaling using Spark

Conclusion

This project demonstrates the power of Big Data analytics in retail. By combining scalable technologies like Spark with machine learning models, we can uncover actionable insights into customer behavior and build intelligent systems for product recommendations and reordering prediction.