Market Basket Analysis using Apriori Algorithm

Harshil Mukesh Bhavsar  
Artificial Intelligence and Machine Learning  
Lambton College In TorontoToronto, Canada  
[c0863170@mylambton.ca](mailto:c0863170@mylambton.ca)

Ritu Vinodbhai Kumbhani  
Artificial Intelligence and Machine Learning  
Lambton College In TorontoToronto, Canada  
[c0865677@mylambton.ca](mailto:c0865677@mylambton.ca)

Shivam Arvindbhai Machhi  
Artificial Intelligence and Machine Learning  
Lambton College In TorontoToronto, Canada  
[c0871186@mylambton.ca](mailto:c0871186@mylambton.ca)

Jeet Dhaneshkumar Patel  
Artificial Intelligence and Machine Learning  
Lambton College In TorontoToronto, Canada  
[c0868588@mylambton.ca](mailto:c0868588@mylambton.ca)

Abstract — Market basket analysis is a widely used data mining technique to identify patterns of co-occurrence among products that are frequently purchased together by customers. The Apriori algorithm is a popular method for performing market basket analysis, which works by iteratively generating frequent itemsets and association rules. This paper presents a proposal of market basket analysis with the Apriori algorithm, including its key concepts, methodologies, and practical applications. We discuss how the algorithm can be used to identify relevant product combinations and optimize product placement and promotion strategies. Additionally, we highlight the challenges and limitations of Apriori algorithm and offer insights into its future directions. Overall, this paper explains the purpose behind the market basket analysis and the Apriori algorithm as a valuable tool for improving business decision-making in the retail industry.

Keywords—machine learning, apriori algorithm, market basket analysis, data mining, pyspark, sqlcontext, pandas, seaborn

# Introduction

In this digital era, shopping on internet has grown largely popular among people, and it's delicate to find someone who has not been exposed to it. Let’s look at an introductory situation in one’s original grocery store to help with understanding what this paper explains about. For instance, if a store has increased chuck sales then it can further upsell it by lowering the price of adulation and jam, as a customer who buys chuck is more likely to buy jam and adulation, all together. So, when it comes to gaining client perceptivity, market basket analysis in data mining remains a crucial factor. As a result, market basket analysis in data mining is a fashion for relating useful and important styles of constantly bought products in a store’s sale history [1].

Market basket analysis (MBA) is a data mining technique that allows one to relate to purchase patterns in any retail terrain. MBA is a set of statistical affinity computations that punctuate coping patterns to help business leaders understand – and eventually serve – their guests more. MBA, in its utmost introductory form, quests for the most common product combinations in deals [2]. To state it simply, MBA is a data mining technique that allows a store proprietor to dissect and determine product combinations, which particulars are related, and which particulars guests constantly buy together. It’s a lovely strategy grounded on the introductory principle that if we buy commodity, we ’re obliged to buy or avoid commodity differently (or a bunch of effects).

# Dataset

The dataset which we used here consists of 7 attributes helpful which help to analyze customer buying patterns. The dataset contains 522065 rows. The 7 features are as follows:

**BillNo**: 6-digit number assigned to each transaction. Nominal.

**Itemname**: Product name. Nominal.

**Quantity**: The quantities of each product per transaction. Numeric.

**Date**: The day and time when each transaction was generated. Numeric.

**Price**: Product price. Numeric.

**CustomerID**: 5-digit number assigned to each customer. Nominal.

**Country**: Name of the country where each customer resides. Nominal.

# Technology used

## PySpark

PySpark is an open-source distributed computing framework that provides a simple and user-friendly interface for processing and analyzing large datasets using the Python programming language. It is built on top of the Apache Spark engine, which enables fast and efficient processing of big data. PySpark provides a range of tools and libraries for data processing, including Spark SQL for structured data, Spark Streaming for real-time data, and MLlib for machine learning tasks. It also supports integration with other big data technologies, making it a versatile tool for data scientists and engineers. With its distributed computing capabilities, PySpark allows users to handle big data processing and analytics with ease, making it an essential tool for big data professionals to model building.

## Pandas

Pandas is fast, powerful and easy to use open-source data analysis and an open-source library built on top of Python Programming language. In this approach we have used pandas to store, manipulate and analyze market data. Pandas works well with other modules for easy data analysis and data preprocessing tasks.

## Seaborn

Seaborn is a powerful and popular open-source data visualization library for Python. It is built on top of the matplotlib library and provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is designed to work with pandas data structures and provides a range of advanced visualization tools, including heatmaps, time series plots, regression plots, and categorical plots. Seaborn also provides a range of customization options, making it easy to create high-quality visualizations that are tailored to specific data sets and research questions. With its intuitive and flexible interface, Seaborn is an essential tool for data scientists and researchers looking to explore and visualize complex data sets.

##### 

##### References

1. Apriori Algorithm in Data Mining: Implementation With Examples. (2023, February 17). Software Testing Help. [https://www.softwaretestinghelp.com/apriori-algorithm/](https://www.softwaretestinghelp.com/apriori-algorithm/\)
2. Ganiyu, I. S. (2023, January 18). Market Basket Analysis in Data Mining Simplified 101 - Learn | Hevo. Learn | Hevo. <https://hevodata.com/learn/market-basket-analysis-in-data-mining/>