INDIANA UNIVERSITY BLOOMINGTON

CSCI B 565

DATA MINING

MARKET BASKET ANALYSIS, CUSTOMER SEGMENTATION AND LIFETIME VALUE PREDICTION

*Authors: Professor:*

Akshat Arvind Yuzhen Ye

Aniket Kale

December 13, 2021

Abstract

Dataset Information

Tables present in the dataset -

* **Products** –

**Columns -** [product\_id, product\_name, aisle\_id, department\_id]

**Size –** 813.67 KB

* **Orders** –

**Columns -**[order\_id,user\_id,eval\_set,order\_number,order\_dow,order\_hour\_of\_day days\_since\_prior\_order]

**Size –** 32.81 MB

* **Order\_products\_\_prior** –

**Columns** - [order\_id, product\_id, add\_to\_cart\_order, reordered]

**Size** – 164.69 MB

* **aisles** –

**Columns** - [aisle\_id, aisle\_name]

**Size** – 1.91 KB

* **Department** –

**Columns -** [department\_id, department\_name]

**Size –** 804B

Introduction

Using the market basket data from Kaggle Instacart Market Basket Analysis competition, we have targeted 4 major objectives in our project –

* Exploratory Data Analysis on data to extract actionable insights.
* To come up with a model to predict customer segmentation and the customer lifetime value using RFM Analysis.
* Use Association Rule Mining to give suggestions of possible products which could be added in a customer’s order basket.
* Also, we have tried to complement the prediction results by adding rewards, promos or clubbed offers to specific customer segments using Association Rule Mining to convert this model into a possible profitable business model.

**Methods**

**Objective -1 EDA**

Exploratory Data Analysis has been performed on the dataset to extract insights and trends from the data.

Tools Used: -

* Matplotlib
* Seaborn
* Pandas
* Plotly

**Objective -2 Association Rule Mining**

Association Rule Mining has been done to suggest additional products based on the products present in the customer’s order basket.

Tools Used: -

* mlxtend
* fpgrowth
* association rules

**Objective -3 Customer Segmentation**

**Objective -4 Customer Lifetime Value**

**Results**

**Objective – 1 EDA**

These are some of the interesting insights extracted from the data –

Chart, histogram

Description automatically generated

The plot shows the count of orders vs days since prior order for customers. We can observe the spikes on 7th, 14th, 21st and 30th day confirming the weekly and monthly trend of reordering supplies.

Chart, histogram

Description automatically generated

The plot shows the count of orders vs order hour of the day. The maximum orders are placed during the day hours from 9 AM – 5PM.

Chart, line chart

Description automatically generated

The plot shows the reorder ratio of items from different departments. We can see that item from bakery, supplies and produce have high reorder ratio compared to items like personal care and household which are products that are frequently ordered.

Graphical user interface, application

Description automatically generated with medium confidence

The plot shows the number of items that customers usually keep in their order basket. According to the data, we can see that an average of 5-10 items are added in the basket by customers before placing their orders.

Chart, histogram

Description automatically generated

The plot shows the number of occurrences of a product type across all the orders in the dataset. We can see that fresh fruits and vegetables are highest ordered products and baby food and herbs are least ordered.

A picture containing graphical user interface

Description automatically generated

The plot shows the first and second product that the customers add in their order basket while placing an order. Produce and dairy eggs are added first by majority of the customers.

**Objective – 2 Association Rule Mining**