Project Proposal

Instacart Market Basket Analysis

Prof. Dr. Thomas Lavastida

BUAN 6340 – Programming for Data Science

Analysis Done By

- Kushagra Rastogi (kxr220031)
- Varun Vinod (vxv220005)
- Aditya Kiran Anasane (aka220002)
- Tanya Shahab (txs220007)

Introduction and Motivation

Market Basket Analysis is a data mining technique used in retail settings to identify associations between items. The method involves analyzing combinations of items that appear together frequently in transactions, allowing retailers to identify relationships between the items that customers purchase. In simpler terms, it helps retailers find patterns in customer behavior with respect to item purchases.

The project provides an analysis of customer purchase patterns, frequently purchased items and units from Instacart to maintain sufficient product inventory and facilitate reordering. The team aims to identify customer clusters and subgroups with comparable purchasing behaviors to provide useful recommendations and improve revenue and customer experience via segmentation and prediction models.

Furthermore, we'll try to predict whether a product will be reordered or not by using order-related and time-related features from the dataset. The results and insights from the analysis will be utilized to enhance user experience by recommending likely products for customers to purchase during the order process. Additionally, a marketing strategy for Instacart and other similar retailers will be proposed, including customized communications, that highlight anticipated products to remind customers to reorder. In conclusion, this project offers Instacart valuable insights to enhance customer experience and increase revenue by employing efficient segmentation and prediction models.

Data Description

The Instacart Market Basket Analysis Kaggle dataset consists of 6 different datasets that provide transactional and purchasing details of customers. The first two datasets, Aisles and Departments, provide information on the names and IDs of the aisles and departments where products were organized. The third and fourth datasets, Order_Products_prior and Order_Products_train, provide information on the orders, products, and reordered products. The fifth data set, Orders, provides information about customer orders such as order ID, order number, weekday of the order, an hour of the order, user ID, and days since the prior order. The sixth dataset, Products, gives information on the products such as product name, product ID, aisle, and departments where they were sold. The objective of this project is to predict the next products customers tend to purchase and whether a product is reordered or not in their next purchase.

A few details about the dataset are as follows:

- No. of Rows: 33819106
- No. of Columns: 15

Some of the key data points required for the analysis in this project are: Transactional details of the products, order number, date and time of the transactions, aisles and departments where the product belongs, Reordered details, time and day details of the products ordered, transactional details, departments and aisles details.

Table: Aisles

Variables	Description
Aisle ID	Labels the ID of the aisles
Aisle Name	Mentions the aisle name in the retail stores

Table: Department

Variables	Description
Department ID	Labels the ID of the departments
Department name	Mentions the department name in the retail stores

Table: Order_Products_prior

Variables	Description
Order_id	Id of the order
Product_id	Id of the products
add_to_cart_order	The order which is added to the cart
reordered	The product that re-order

Table: Order_Products_train

Variables	Description
Order_id	Id of the order
Product_id	Id of the products
add_to_cart_order	The order which is added to the cart
reordered	The product that re-order

Table: Orders

Variables	Description
Order ID	Labels the ID of the order made by customers
User ID	Labels the ID of the users who made the purchase
Order number	Denotes the order number made by the customer
Order_dow	Denotes the day of the week, the order made by the customer
Order hour of day	Denotes the hour of the day, the order made by the customer
Days since prior order	Number of days since the last order

Table: Products

Variables	Description
Product ID	Labels the ID of the products purchased by customers
Product Name	Denotes the product name purchased by the customer
Aisle ID	Labels the ID of the aisles
Departments ID	Labels the ID of the departments

Project Outline

1. **To create clusters of users based on everyday purchase habits.**

   The Instacart dataset provides us with User Ids and their respective order IDs which correspond to a certain product. Products are also grouped as per the aisles.

   Based on this data we can create groups/clusters of userIds (users) who have similar product purchasing habits to identify the different customer segments, analyze the clusters and unearth consumer behavioral patterns, and basis these patterns provide suggestions and recommendations to Instacart in order to drive sales, revenue, and customer traffic.

2. **To predict if a product will be reordered or not.**

   Instacart provided data on the past orders of 200,000 users, which are classified as prior, train, or test orders. The goal is to predict which products will be in a user's future demand, which is a classification problem. Predictor variables (X) will be calculated based on the characteristics of the product and the user's behavior. The Logistic regression algorithm is used to create a predictive model, which is applied to the test dataset to predict the "reordered" variable. The method includes importing features, creating test and training Data Frames, creating the predictive model, and applying the model.

References

Kaggle: https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis?select=departments.csv