**Instacart Market Basket Analysis**

**By Stacey Baptiste**

**SDC-486LAB**

**A colorful logo with a white background

AI-generated content may be incorrect.**

**Analytical Report**

**Title and Overview:**

Instacart is a leading online grocery delivery service that has revolutionized the way consumers shop for groceries. With our growing population understanding customer spending practices is essential for a competitive edge. This project focuses on examining customer purchasing behavior utilizing Instacart transaction dataset. By utilizing data analytics, I was able to identify key patterns, refine marketing strategies, and enhance product recommendations. Through this research data mining techniques were performed through machine learning. This discovered frequently purchased products; segment of customers purchasing behavior and enhance recommendations and target marketing strategies for retailers.

**Problem Statement:**

In a widely competitive grocery retail market, offering promotions, discounts and recommendations of products enhances customers’ shopping experiences. These strategies help build strong relationships, better understand customers’ needs, and effectively target loyal shoppers. One key challenge in this area is to identify customer segmentation based on purchase history. Customers were grouped into distinct segments based on their purchasing patterns. Additionally, understanding temporal trends in purchasing behavior is crucial. How do buying patterns fluctuate over time, daily, weekly, or seasonally? Are certain products more popular at specific times of the year or day?

***Solution:***

By addressing these questions enables businesses to optimize staffing, inventory, and marketing efforts while allowing for targeted promotions and personalized recommendations, ultimately improving customer satisfaction and operational efficiency. The characteristics defined customers’ purchasing behavior by hour of the day orders was placed, reordered items placed, number of orders, and the average days since the prior orders.

**Objectives:**

* I was able to analyze customer purchasing habits/patterns to (eg: reordered items, days\_since\_prior\_order, order\_hour\_of\_day, product\_name, aisle, department) to optimize inventory and target promotions.
* Identified product frequency, bundle deals, and recommendation strategies
* Segment customers based on their purchasing behavior to personalize promotions and discounts.

**Data Description:**

* The data for this project is the Instacart Online Grocery Shopping dataset [Kaggle Instacart Dataset](https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data?select=order_products__train.csv) which contains over 3 million grocery orders from more than 200,000 Instacart users. The dataset includes the following:
  + - Aisles (aisle\_id, aisle)
    - Departments (department\_id, department)
    - Order Product Prior (order\_ID, product\_ID, add\_to\_cart\_order, reordered)
    - Order Product Train (order\_ID, product\_ID, add\_to\_cart\_order, reordered)
    - Orders (order\_ID, user ID, eval\_set, order\_number, order\_dow, order\_hour\_of\_day, days\_since\_prior)
    - Products (product ID, product name, aisle\_ID, department\_ID)

**Data Cleaning**

The Instacart dataset was retrieved from [Kaggle Instacart Dataset](https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data?select=order_products__train.csv). After downloading and removing unnecessary columns the following data cleaning process was addressed: Identified missing values filling them with the mode to ensure no data was missing while maintaining consistency with the most common value. Duplicates and inconsistent entries were carefully removed and displayed missing values after cleaning.

**Preprocessing Steps**

Categorical variables (product name, aisle, and department) were encoded using one-hot encoding, and features were scaled using standardization. Feature engineering was performed to predict repeat purchases for optimizing recommendations and inventory planning. The dataset was split into training and testing sets with a ratio of 80:20. Data transformation techniques were applied as needed to address specific issues and prepared the data for model training.

**Exploratory Data Analysis (EDA)**

**Python libraries used:**

* import pandas as pd
* from IPython.display import display, Markdown
* import matplotlib.pyplot as plt
* import seaborn as sns
* from sklearn.preprocessing import OneHotEncoder
* from sklearn.preprocessing import StandardScaler
* import tensorflow as tf
* import numpy as np
* import matplotlib.pyplot as plt
* from tensorflow.keras.models import Sequential
* from tensorflow.keras.layers import Dense, Dropout
* from sklearn.preprocessing import StandardScaler
* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import classification\_report, mean\_squared\_error, silhouette\_score, precision\_score, recall\_score, f1\_score, confusion\_matrix
* from sklearn.ensemble import RandomForestClassifier
* from sklearn.metrics import accuracy\_score, classification\_report
* import tensorflow as tf
* from tensorflow import keras
* from tensorflow.keras.models import Sequential
* from tensorflow.keras.layers import Dense
* from tensorflow.keras.optimizers import Adam
* import matplotlib.pyplot as plt

Key feature distributions were analyzed using histograms, pair plot, heatmap, and box plots to examine central tendencies and data dispersion. This helped uncover underlying patterns and detect potential anomalies. A correlation matrix was generated to investigate relationships between variables, with heatmaps providing a visual representation of significant correlations and possible confounding factors. Additionally, scatter plots were utilized to identify trends and patterns within the dataset. These visualizations helped understand the nature of relationships between variables and identify any potential non-linear trends.

* Visualize key statistics using Python (Pandas, Matplotlib, Tensorflow) and PowerBI.

***Univariate Analysis:***

* Performed a summary statistic to understand the central tendency and dispersion of numerical features
* Created histograms visualizations to show the distribution of numerical features.
* Used box plots to detect and visualize outliers in the numerical features.

***Bivariate/Multivariate Analysis:***

* Constructed a correlation matrix to identify relationships between numerical variables
* Generated a set of pair plots to visualize pairwise relationships between features
* Created a heatmap to visualize complex relationships and correlations across the dataset
* Created histograms to visualize the distribution of the numerical features
* Used box plots to detect and visualize outliers in the numerical features

**Baseline Model Statistics**

Linear Regression was selected as the baseline model for its simplicity and ease of interpretation. While alternative models such as Decision Trees and Linear Regression were considered, Random Forests served as a strong initial benchmark. The model's performance was assessed using metrics including accuracy score and classification report offering a quantitative measure of its predictive capabilities. Training and validation scores were examined to evaluate the model's generalization to unseen data. Additionally, a confusion matrix and classification report were generated to analyze its performance in classification tasks.

***Baseline Predictive Model:***

The baseline model, Random Forest, was implemented by splitting the preprocessed data into training and test sets. Its performance was evaluated using accuracy and regression metrics, identifying areas for improvement in preparation for neural network development.

***Neural Network Implementation:***

Designed the neural network architecture, defining layers, neurons, and activation functions. Trained the model on the training data while monitoring test performance. Evaluated accuracy and other relevant metrics, such as precision, recall, or mean squared error. Analyzed strengths and weaknesses using visualizations like confusion matrices, identifying areas for improvement, including underfitting, overfitting, and hyperparameter tuning.

**Scenario Analysis**

Conducted scenario analysis to assess the impact of key performance indicators (KPIs) of customers purchase patterns. I identified three scenarios by key features to explore various outcomes: Items Reordered by Aisle, Average of reordered by department, and the Sum of days since prior order. The neural network model was used to predict outcomes for each scenario and compared results to baseline predictions. Next, I Analyzed how changes influenced model outputs giving a general sense of customer ordering habits.

**Data visualization**

Developed an interactive Power BI dashboard to visualize key insights from previous project phases. Prepared a layout of critical analysis results and selected appropriate visualizations, such as bar charts, line graphs, and heatmaps. Integrated EDA findings, predictive model metrics, and neural network results, highlighting key comparisons baseline and neural network performance. Visualizations identified feature of importance and relationships, along with dynamic elements like filters and slicers.

**References:**

* *Instacart Market basket analysis*. (2017, November 20). Kaggle. <https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis/data?select=order_products__train.csv>
* Mining techniques: Kanade, V. (2022, October 4). *Data Mining: Definition, techniques, and tools - Spiceworks Inc*. Spiceworks Inc. <https://www.spiceworks.com/tech/big-data/articles/what-is-data-mining/>
* GeeksforGeeks. (2025, January 15). K means Clustering Introduction. GeeksforGeeks. <https://www.geeksforgeeks.org/k-means-clustering-introduction/>
* Analytical-Report-Building-Samples. (n.d.). In *Gamma.com*. <https://Analytical-Report-Building-Samples.pdf>
* Data visualization tools (Matplotlib, TensorFlow, Power BI):
  + GeeksforGeeks. (2025b, January 16). *Data Visualization using Matplotlib in Python*. GeeksforGeeks.<https://www.geeksforgeeks.org/data-visualization-using-matplotlib/>
  + *W3Schools.com*. (n.d.-d). <https://www.w3schools.com/ai/ai_tensorflow_intro.asp>
  + Kenji Explains. (2023, March 5). Master Power BI Essentials in Just 15 minutes [Video]. YouTube. <https://www.youtube.com/watch?v=nkmHqs1I_z0>