ISYE 6740 Course Project

Sidarth Sudhakar

June 30, 2024

Abstract

This project aims to leverage Instacart's extensive dataset to predict which previously purchased products will be included in a user's next order. By employing advanced feature engineering and machine learning techniques, we will develop a classification model to enhance Instacart's recommendation system. This will not only improve customer satisfaction by anticipating their needs but also drive sales through personalized shopping experiences.

1 Introduction

The rapid growth of the quick-commerce segment, projected to reach $3.1 billion by 2025, has created a massive influx of data, especially in the grocery delivery sector. Instacart, a leading grocery ordering and delivery platform in the United States, aims to capitalize on this data to revolutionize consumer shopping habits. By predicting user buying patterns and optimizing order value, Instacart can enhance the user experience and increase repeat purchases. This project focuses on predicting whether a user will reorder previously purchased products, thereby offering a personalized and efficient shopping experience.

2 Dataset

Instacart has provided a comprehensive dataset comprising orders from 200,000 users, each having between 4 and 100 orders. The data is categorized into prior, train, and test orders. Prior orders reflect past user behavior, while train and test orders pertain to future behavior. The dataset includes 3.4 million orders, with 3.2 million available for feature engineering from prior orders. There are approximately 50,000 products across 21 departments, stored in 134 aisles. Our task is to predict the products in a user's next order using this rich dataset. The ERD of the dataset is represented in Figure 1 below.

3 Methodology

For the user to get product recommendations based on his past orders, the model should be able to learn from the patterns and generated probability against each product which will indicate chances of re-ordering.

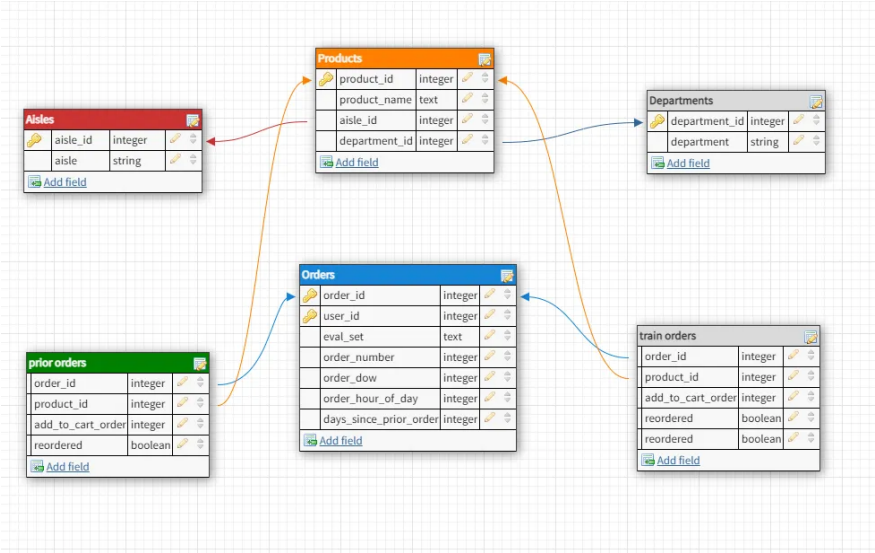
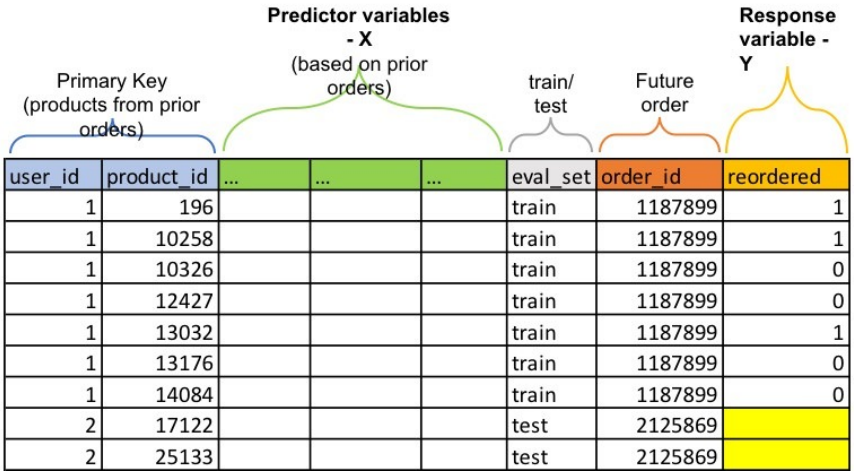


Figure 1: Data Schema

Apart from this, upon learning from other users, new product recommendation should be provided to the users with higher probability of getting added to the cart. Additionally, the cold start scenario [New user], should also be addressed with baseline predictions.



user_id	product_id	...	...	...	eval_set	order_id	reordered
1	196				train	1187899	1
1	10258				train	1187899	1
1	10326				train	1187899	0
1	12427				train	1187899	0
1	13032				train	1187899	1
1	13176				train	1187899	0
1	14084				train	1187899	0
2	17122				test	2125869	
2	25133				test	2125869	

Figure 2: Problem Statement

On a high level, the dataset would be merged and cleaned. Using EDA, features have to be generated against each user-product index. This dataset would then be used to predict the re-ordered variable along with a probability. Additionally, association mining rules would be used to suggest new products upon learning patterns from all the other users as well. The same is represented diagrammatically in Figure 2.

3.1 Exploratory Data Analysis

1. Data Cleaning and Preprocessing

(a) Address missing values and outliers

(b) Convert categorical variables into a usable format

(c) Understand the impact of other-meta product information to aid in feature engineering

2. Feature Engineering

   (a) **User Segmentation**: Classify users based on purchasing patterns using PCA and kMeans.

   (b) **User-Product-Order Features**: Generate features capturing user interaction with products

   (c) **Product Features**: Utilize techniques like Word2Vec to create product embedding based on textual descriptions.

   (d) **Meta information**: Day and time of item purchased by the user.

3. Modeling The approach would mainly be an ensemble modeling to handle different scenarios. The main focus would be to generate a probabilistic appraoch and then pick the top products based on threshold.

   (a) **XGBoost**: A powerful gradient boosting framework known for its accuracy and efficiency.

   (b) **LightGBM**: A gradient boosting framework that handles large datasets and complex features effectively.

   (c) **Apriori Algorithm**: Identify frequent item sets and association rules.

   (d) **Markov Chains**: Model sequential purchasing behavior and predict next-order products.

4 Validation Strategy

By accurately predicting reorder patterns, Instacart can enhance its recommendation engine, leading to increased user satisfaction and loyalty. This personalized approach not only boosts sales but also streamlines the shopping process, making it more intuitive and efficient for users. Implementing these models will provide actionable insights into customer behavior, allowing Instacart to tailor its services and marketing strategies effectively.

1. **F1-Score**: Balance between precision and recall, ensuring customer satisfaction and a smooth user experience. This would be helpful in gauging the baseline performance.

2. **F1 Score Maximization**: Instead of applying global threshold, a local threshold could be applied for each output which accounts for the combination of products.

3. **ROC-AUC Curve**: Evaluate the model's performance across different threshold levels, optimizing for the right tradeoff between sensitivity and specificity.

5 References

1. F1-Score Maximisation https://icml.cc/Conferences/2012/papers/175.pdf
2. Dataset https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2
3. Handling large dataframes https://towardsdatascience.com/make-working-with-large-dataframes-easier-at-least-for-your-memory-6f52b5f4b5c4
4. Ensemble Modeling using scikit-learn https://towardsdatascience.com/ensemble-learning-using-scikit-learn-85c4531ff86a