Feature Engineering Report

Ankita Upadhyay

December 3, 2024

1 Introduction

This report details the various features constructed during the feature engineering phase for the recommendation system.

2 Feature Engineering

2.1 Time of Day Buckets

Time of day is categorised into four buckets: morning, afternoon, evening, and night. This categorisation helps capture temporal patterns in user behavior.

2.2 Days Since Last Purchase

This feature calculates the number of days since a user last purchased a specific product. It provides valuable insights into a user's purchase cycle.

2.3 Number of Orders Per User

This feature tracks the total number of orders placed by each user. Users who place more orders tend to be more engaged with the platform.

2.4 Sum of Items Ordered Per Order

This feature calculates the total number of items ordered in each individual order. Orders with a larger number of items may suggest bulk buying.

2.5 Total Orders for Each Product

The total number of orders for each product is calculated to measure its popularity. Products that are frequently ordered by many users should be prioritized in product recommendations.

2.6 Average Order Hour for Each User

This feature calculates the average hour of the day when a user places orders.

2.7 Order Frequency

Order frequency tracks how often a user purchases the same product. Products that are frequently ordered by a user indicate strong preferences.

2.8 Product Affinity by Department

This feature calculates the proportion of orders placed within each department for each user.

2.9 Time of Purchase Features: Preferred Hour and Day of Week

These features capture the user's preferred time of day and day of the week for placing orders.

2.10 Average Reorder Rate

The average reorder rate measures the likelihood that a user will reorder an item they have previously purchased.

2.11 Product Reorder Frequency (Across All Users)

This feature calculates how often a product is reordered by all users.

2.12 User's Reorder Rate for Product

This feature measures the likelihood that a specific user will reorder a specific product.

2.13 Order Count for Product by User

This feature tracks the number of times a user has ordered a specific product.

3 Data Columns

The following columns were obtained after the feature construction phase:

- product_id, order_id, add_to_cart_order, reordered, user_id, order_number, order_dow, order_hour_of_day
- days_since_prior_order, time_of_day, days_since_last_purchase, user_order_count, total_items_ordered

- product_popularity, avg_order_hour, order_frequency, department_affinity, preferred_order_hour
- preferred_order_dow, avg_reorder_rate, product_reorder_frequency, user_product_reorder_rate, user_product_order_count

4 Feature Scaling

The features were scaled using the Min-Max Scaler to bring them to the same scale. This is important for the Logistic Regression model fitting.

5 Dataset Distribution After Scaling

The distribution of the dataset after performing Min-Max Scaling is as follows:

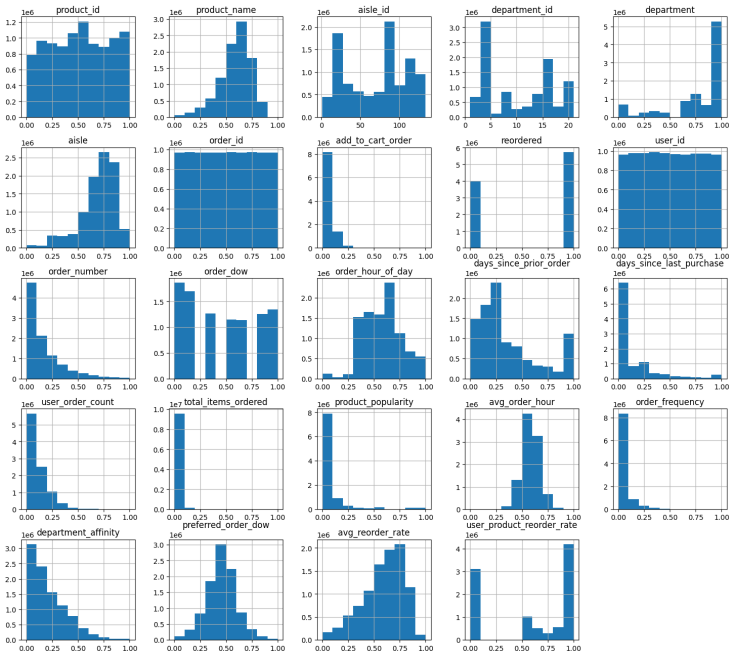

Figure 1: Distribution of dataset after scaling

6 Correlation Analysis

Here is the correlation matrix that shows the relationships between the features:

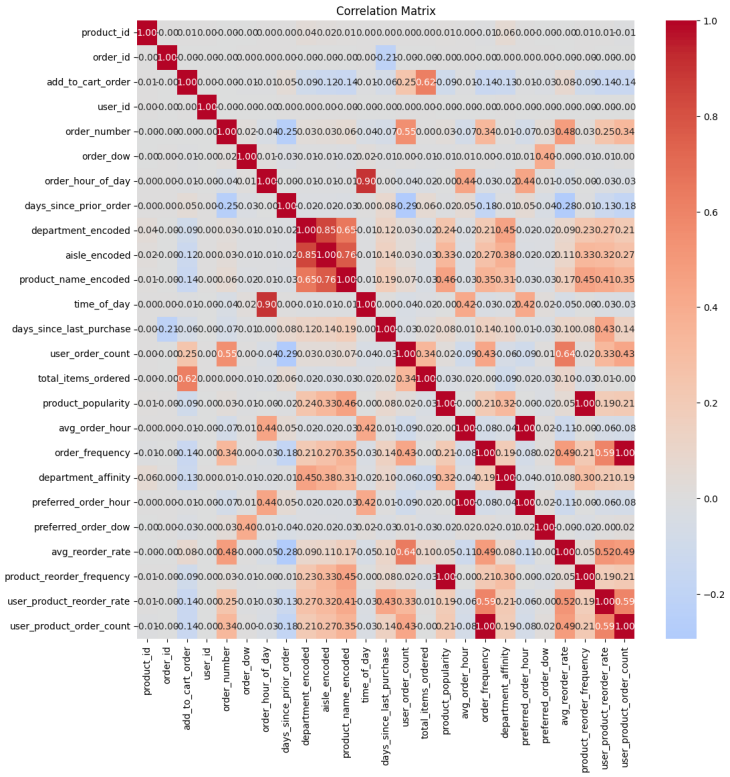

Figure 2: Correlation matrix of the features

7 Feature Importance

After performing feature importance, the most important features for model fitting were identified. These features are visualized in the figure below:

8 ANOVA F-test

The top 10 features selected using the ANOVA F-test are:

- add_to_cart_order, order_number, days_since_last_purchase
- user_order_count, product_popularity, avg_order_hour
- order_frequency, department_affinity, avg_reorder_rate, user_product_reorder_rate

9 Lasso Regularization

The selected features using Lasso regularization are:

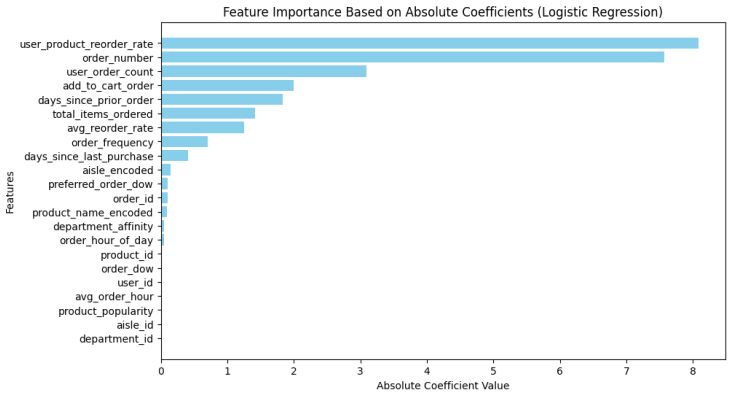

Figure 3: Feature Importance Plot

- order_number, days_since_prior_order, user_product_reorder_rate

10 Model Accuracy

The Logistic Regression model accuracy using F1-score method is 0.9518.