Feature Engineering Report

Ankita Upadhyay

December 4, 2024

1 Feature Engineering Descriptions

1.1 Time of Day Buckets

Time of day is categorized into four buckets: morning, afternoon, evening, and night. This categorization helps capture temporal patterns in user behavior. For instance, some users may prefer ordering in the morning, while others might prefer evening or night orders. This feature is useful for predicting which products a user may be interested in purchasing at different times of the day, aligning recommendations with the user's preferred order times.

1.2 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 and can help identify products that may be due for reordering. By incorporating this feature, the system can better predict products that the user might be interested in purchasing again, improving the likelihood of timely and relevant recommendations.

1.3 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 and may have a higher likelihood of purchasing similar items in the future. By understanding user activity and engagement through this feature, we can better personalize product recommendations based on individual user behavior.

1.4 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 or more diverse product preferences. This feature helps in identifying products that are frequently bought together, as well as in understanding a user's overall purchasing habits, which in turn improves the recommendation system by suggesting items aligned with user preferences.

1.5 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, as they have a higher likelihood of being purchased again. This feature helps identify trending or highly sought-after products that are good candidates for recommendation to users.

1.6 Average Order Hour for Each User

This feature calculates the average hour of the day when a user places orders. By understanding a user's preferred time for ordering, the recommendation system can align product suggestions with the user's natural shopping rhythm. Users who typically order at specific times of the day may be more likely to engage with products recommended during those hours, improving conversion rates.

1.7 Order Frequency

Order frequency tracks how often a user purchases the same product. Products that are frequently ordered by a user indicate strong preferences or habits. This feature is crucial for identifying repeat purchases and cross-selling opportunities. By incorporating order frequency, the system can suggest products that a user is likely to reorder based on their past buying behavior.

1.8 Product Affinity by Department

This feature calculates the proportion of orders placed within each department for each user. It helps to determine a user's affinity for specific product categories or departments (e.g., groceries, electronics). Understanding a user's department preferences enables the recommendation system to offer personalized product suggestions within the user's favourite categories, enhancing the relevance of the recommendations.

1.9 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. By analyzing when a user typically makes purchases, the system can recommend products at optimal times, increasing the likelihood of a purchase. This temporal personalization allows the recommendation engine to suggest products based on the user's habitual shopping times, improving the overall user experience.

1.10 Average Reorder Rate

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

user tends to repurchase items, which is valuable for recommending products that they may wish to reorder. This feature enhances the system's ability to suggest relevant items based on a user's past purchase history.

1.11 Product Reorder Frequency (Across All Users)

This feature calculates how often a product is reordered by all users, not just a specific individual. Products with a high reorder frequency are good candidates for recommendation, as they have a strong likelihood of being bought again by other users. This feature helps surface popular items that are likely to be of interest to many users, based on their tendency for repeat purchases.

1.12 User's Reorder Rate for Product

This feature measures the likelihood that a specific user will reorder a specific product. It captures individual user-product interactions, which are critical for personalized recommendations. By including this feature, the system can better predict which products a user is most likely to reorder, improving the personalization of product suggestions.

1.13 Order Count for Product by User

This feature tracks the number of times a user has ordered a specific product. It helps identify products with high user-specific engagement, which can be used to create more personalized recommendations. Products that a user has ordered more frequently are likely to be recommended again, especially if they align with the user's preferences and past purchasing behavior.

2 Columns After Feature Construction

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'

3 Correlation Matrix

The correlation matrix of different created features are as follow

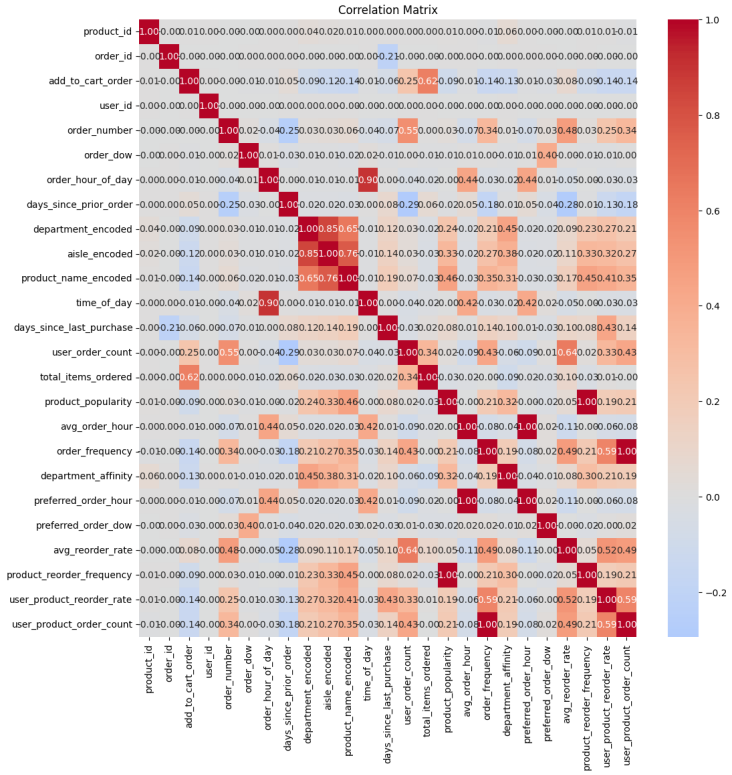

Figure 1: Confusion Matrix of the Logistic Regression Model

4 Feature Importance

Upon performing the feature importance step, the following features were found to be the most important for model fitting:

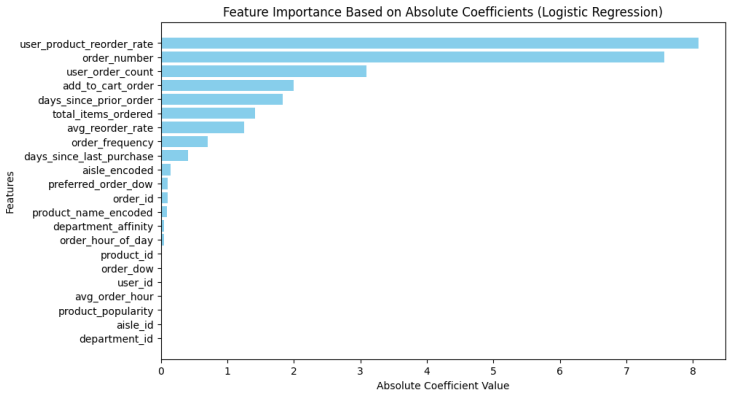

Figure 2: Feature Importance Bar Chart

5 ANOVA F-Test

The following top 10 features were selected based on the ANOVA F-test:

Top 10 features based on ANOVA F-test: $['add\_to\_cart\_order', 'order\_number', 'days\_since\_last\_purchase', '$

These features can be further used for fitting the model and improving results.

6 Selected Features Using Lasso Regularization

The features selected using Lasso regularization are as follows:

Selected features using Lasso regularization: $['order\_number', 'days\_since\_prior\_order', 'user\_product\_reorde$

7 Scaling

Scaling was applied using Min-Max Scaler to bring all features to the same scale, making it suitable for the Logistic Regression function.

8 Distribution of the Dataset After Min-Max Scaling

The distribution of the dataset after applying Min-Max scaling is shown below:

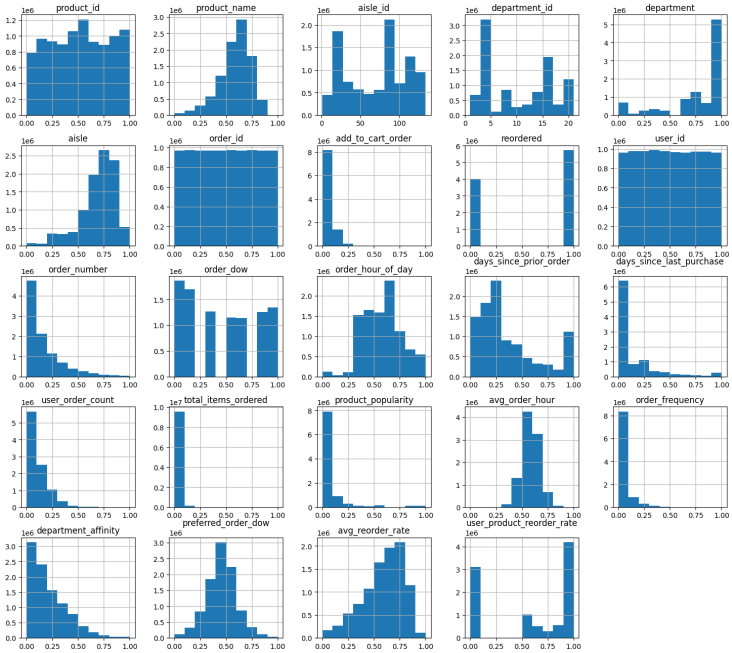

Figure 3: Distribution of Scaled Dataset

8.1 Insights from Histograms

- **product_id:** Distribution of product usage across orders. Popular products will have higher frequencies.
- **product_name:** Reflects the popularity distribution of products based on frequency.
- **aisle_id and aisle:** Identifies frequently visited aisles. Skewness could suggest some aisles dominate in orders.
- **department_id and department:** Highlights which departments (e.g., produce, dairy) contribute most to orders. Uneven distributions suggest product preferences by department.
- **order_id:** Uniformity or clustering in order IDs might represent data segmentation.
- **add_to_cart_order:** Shows product ranking within the cart. Lower values indicate products that are added earlier, which might be staples.
- **reordered:** Indicates the percentage of reorders vs. new orders. A high frequency at "1" suggests users reorder frequently.

- **user_id:** Helps to visualize user distribution in the dataset. Can identify active users or those making multiple orders.
- **order_number:** Represents the number of orders users have placed. Peaks might indicate patterns in user activity (e.g., loyal customers).
- **order_dow (Day of the Week):** Highlights peak shopping days. Can provide insights into user shopping behavior trends.
- **order_hour_of_day:** Captures popular times for shopping. Peaks can help optimize marketing campaigns or operations.
- **days_since_prior_order:** Time gap between consecutive orders. Peaks can highlight habitual ordering behavior (e.g., weekly, bi-weekly).
- **user_order_count:** Indicates activity levels of users. A small group of users might contribute disproportionately to orders.
- **total_items_ordered:** Total items ordered in each transaction. Peaks can identify average order sizes.
- **product_popularity:** Frequency of products being purchased. Highlights the most in-demand products.
- **avg_order_hour:** Average shopping time for users. Can identify if users prefer morning, afternoon, or evening shopping.
- **order_frequency:** Frequency of product purchases by user-product pairs. Indicates strong product affinities.
- **department_affinity:** Preference for departments by users. Higher values indicate specific department dominance.
- **preferred_order_dow:** User-specific preferences for shopping days. Useful for user-level personalization.
- **avg_reorder_rate:** User tendency to reorder products. High reorder rates suggest user satisfaction with previously purchased items.
- **product_reorder_frequency:** Reorder frequency across all users for each product. Higher values indicate staple items.
- **user_product_reorder_rate:** User-specific tendency to reorder specific products. Can identify strong product loyalty or dependency.
- **user_product_order_count:** Total orders of specific products by users. Reflects the depth of user-product interaction.

9 Model Accuracy

The Logistic Regression model was used to evaluate the accuracy after performing the feature engineering steps. The accuracy of the model, based on the F1-score method, is:

Model Accuracy (using F1-score method): 0.9518

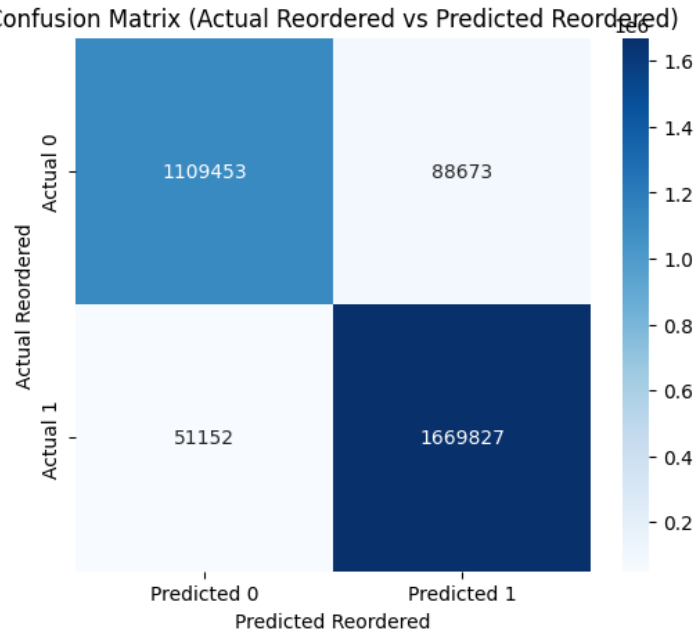

Figure 4: Confusion Matrix of the Logistic Regression Model