Feature Engineering Report

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. 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.

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.

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.

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.

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.

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.
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.

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 personalised product suggestions within the user’s favourite categories, enhancing the relevance of the recommendations.

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

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.

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.

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 personalised recommendations. By including this feature, the system can better predict which products a user is most likely to reorder, improving the personalisation of product suggestions.

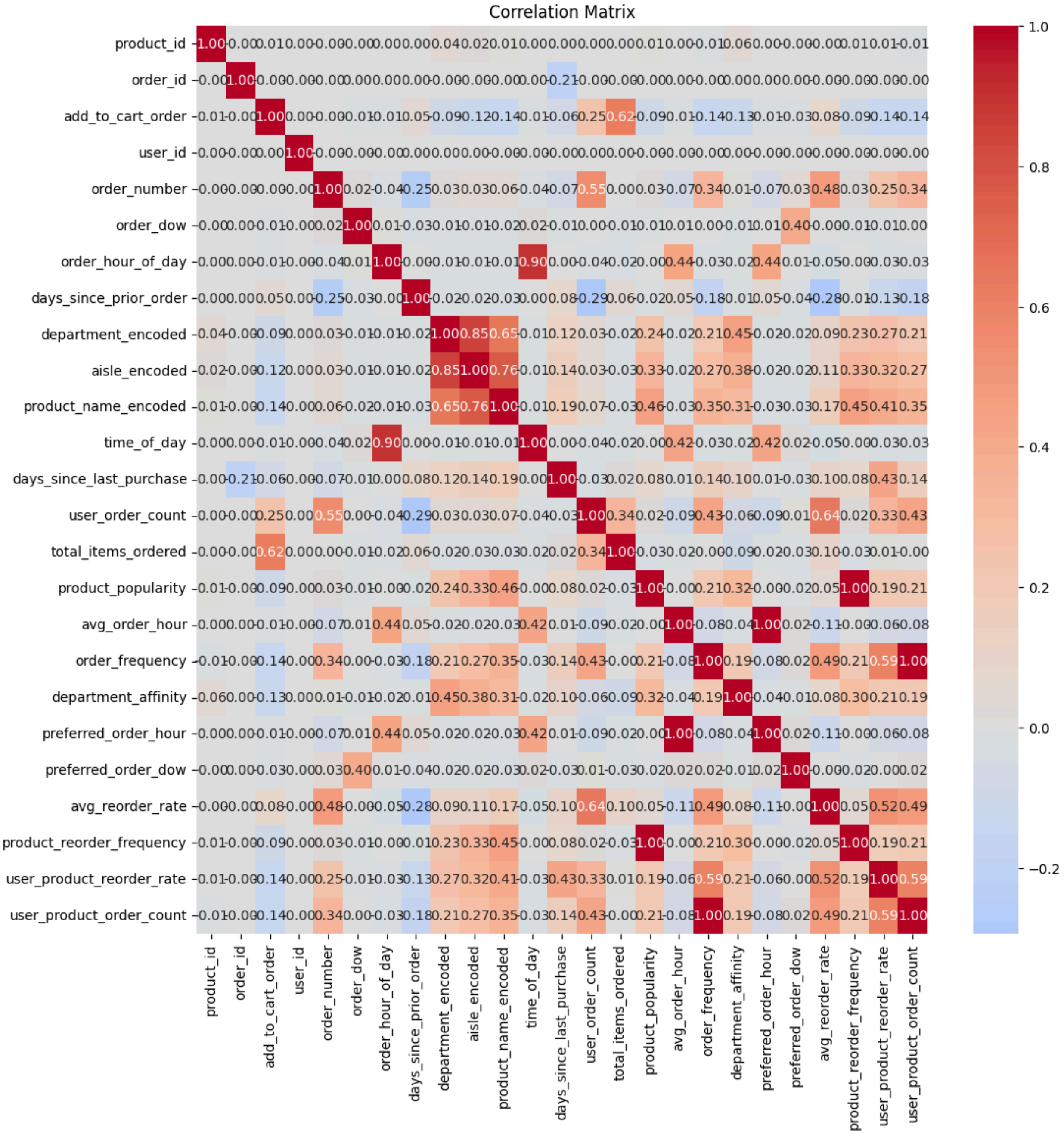
13. Order Count for Product by User
This feature tracks the number of times a user has ordered a specific product. It helps identify products with high user-specific engagement, which can be used to create more personalised recommendations. Products that a user has ordered more frequently are likely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

Above are the descriptions of the features constructed.

```
'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'
```

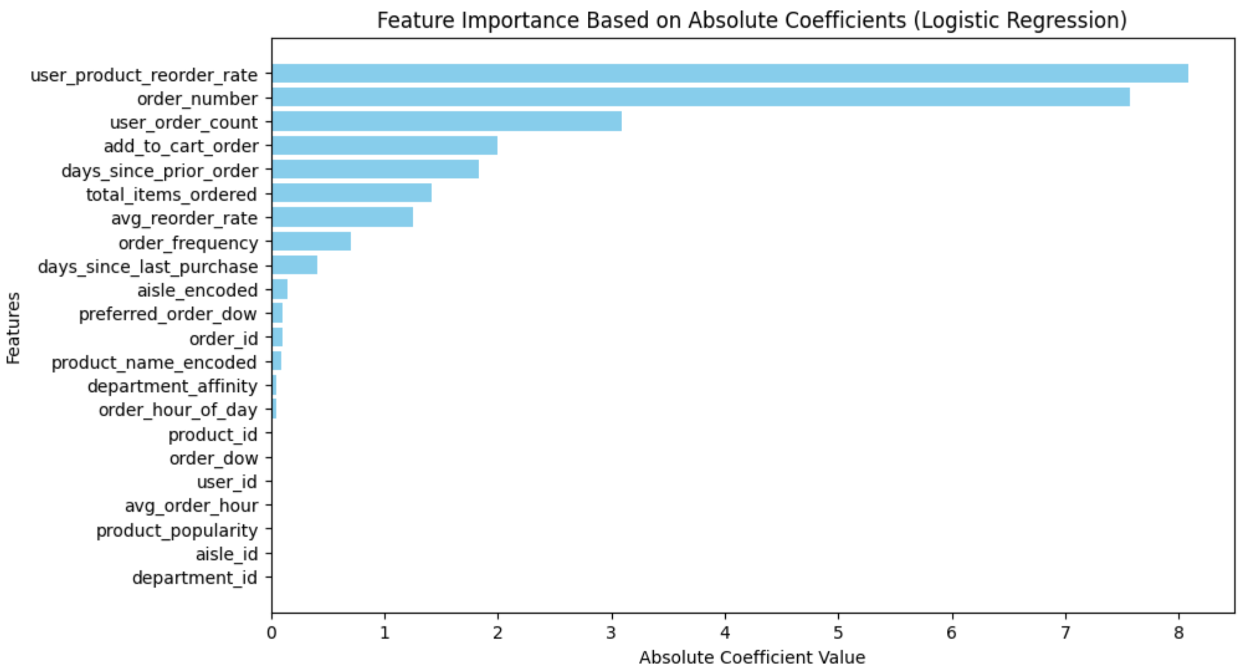
These are the columns obtained after the feature-construction phase.

I have done scaling to bring the features to same scale as it would help in fitting in Logistic Regression function.I have used Min-max Scaler for scaling the dataframe.
Here is the correlation image:



Upon performing the feature importance step I found that these are the important features which can be used for model fitting phase:

Here is the image of feature importance phase:



I have performed the test for selection of ANOVA F-test for selecting features and these are the best features which are obtained:

```
Top 10 features based on ANOVA F-test: Index(['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']
```

These features can be further used for fitting the model and it helps in giving better results.

Selected features using Lasso Regularization is obtained as follow:

```
Selected features using Lasso regularization: Index(['order_number',
'days_since_prior_order', 'user_product_reorder_rate']
```

The Logistic Regression model was used for finding the accuracy of the model after performing feature engineering step.

Given below is the accuracy of the logistic regression model :
Model Accuracy (using f1-score method): 0.8585