Data Encoding and Feature Engineering Report

Prepared by: Priyam Pal

05th December 2024

Contents

1 Data Encoding

1.1 Target Encoding of `aisle`, `department`, `product_name`

What is Target Encoding: Target encoding replaces categorical values with the mean of the target variable (e.g., reordered) for each category.
Features Being Encoded: `aisle`, `department`, `product_name`
Why This Encoding is Needed: Captures the relationship between categorical features and the target variable.

Table 1: Target Encoding Results

aisle_target_enc	department_target_enc	product_name_target_enc
0.547057	0.574195	0.585799
0.590163	0.574195	0.629367
0.671338	0.628133	0.632911
0.469460	0.346935	0.437500
0.493279	0.541577	0.509413

1.2 One-Hot Encoding of `day_of_week`

What is One-Hot Encoding: Converts categorical values into binary columns.
Features Being Encoded: `order_dow`
Why This Encoding is Needed: Allows the model to treat each category as independent.

Table 2: One-Hot Encoding Results

dow_0	dow_1	dow_2	dow_3	dow_4	dow_5	dow_6
0	1	0	0	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1
1	0	0	0	0	0	0

1.3 Cyclic Encoding and Sin-Cosine Transformation

What is Cyclic Encoding: Transforms cyclic features into sine and cosine components.
Features Being Encoded: `order_hour_of_day`
Why This Encoding is Needed: Preserves cyclical relationships.

1.4 Ordinal Encoding of `add_to_cart_order`

What is Ordinal Encoding: Converts ordered data into numerical format.
Features Being Encoded: `add_to_cart_order`
Why This Encoding is Needed: Retains the rank information.

Table 3: Cyclic Encoding Results

order_hour_sin	order_hour_cos
0.258819	-0.965926
0.866025	-0.500000
0.707107	-0.707107
1.22465E-16	-1
-0.965926	-0.258819

Table 4: Ordinal Encoding Results

order_id	add_to_cart_order_encoded
2722718	8
2089674	9
3024155	0
2890872	1
1798802	2

1.5 Binning of `days_since_prior_order`

What is Binning: Categorizes continuous values into discrete intervals.
Features Being Encoded: `days_since_prior_order`
Why This Encoding is Needed: Simplifies continuous data.

Table 5: Binning Results

order_id	days_since_prior_order_binned
2722718	8-15
2089674	24-31
3024155	0-7
2890872	0-7
1798802	Unknown

2 Feature Engineering

2.1 Feature I: Average Days Between Purchases

Definition: Average time interval (in days) between consecutive purchases of a product by a user.
Role: Identifies user buying behavior and product repurchase frequency.

Table 6: Average Days Between Purchases

order_id	product_id	average_days_between_purchases
2722718	8619	15
2089674	13870	29
3024155	2029	4
2890872	16062	2
1798802	14335	7.17

2.2 Feature II: Product Purchase Frequency

Definition: Total number of times a specific product has been purchased by a user.
Role: Highlights user preferences and product loyalty.

Table 7: Product Purchase Frequency

order_id	product_id	product_purchase_frequency
2722718	8619	1
2089674	13870	1
3024155	2029	1
2890872	16062	1
1798802	14335	6

2.3 Feature III: Total Purchases

Definition: Total number of purchases for each user-product combination.
Role: Tracks user engagement with specific products.

Table 8: Total Purchases

order_id	product_id	total_purchases
2722718	8619	1
2089674	13870	1
3024155	2029	1
2890872	16062	1
1798802	14335	6

2.4 Feature IV: Interval Standard Deviation

Definition: Standard deviation of time intervals between consecutive purchases of a product by a user.
Role: Measures variability in purchase frequency.

Table 9: Interval Standard Deviation

order_id	product_id	interval_std_dev
2722718	8619	NaN
2089674	13870	NaN
3024155	2029	NaN
2890872	16062	NaN
1798802	14335	4.45

2.5 Feature V: Product Reorder Rate

Definition: Ratio of the number of times a product has been reordered to its total orders.
Role: Indicates product popularity and customer retention tendencies.

Table 10: Product Reorder Rate

order_id	product_id	product_reorder_rate
2722718	8619	0.585799
2089674	13870	0.629367
3024155	2029	0.632911
2890872	16062	0.437500
1798802	14335	0.509413

2.6 Feature VI: User General Reorder Rate

Definition: Ratio of reordered items to total items purchased by a user.
Role: Captures a user's general tendency to reorder products.

Table 11: User General Reorder Rate

order_id	product_id	user_general_reorder_rate
2722718	8619	0.818182
2089674	13870	0.476190
3024155	2029	0.176471
2890872	16062	0.279070
1798802	14335	0.712500

2.7 Feature VII: Average Add to Cart Order

Definition: Average position of a product in the cart across its purchases.
Role: Reflects product priority in user shopping behavior.

Table 12: Average Add to Cart Order

order_id	product_id	avg_add_to_cart_order
2722718	8619	10.92
2089674	13870	9.49
3024155	2029	9.47
2890872	16062	13.52
1798802	14335	8.59

3 Correlation Matrix

Insights:

- **Strong Correlations:** - Features such as `aisle_target_enc`, `department_target_enc`, and `product_name_target_enc` are highly correlated with `product_reorder_rate`, making them critical predictors of reorder behavior. - `product_purchase_frequency` and `total_purchases` show strong correlation, indicating they represent user-product purchase activity.
- **Negative Correlation:** - `avg_add_to_cart_order` is negatively correlated with `product_reorder_` and `user_general_reorder_rate`, suggesting earlier positions in the cart are linked to higher reorder likelihood.
- **Weak Correlations:** - Features such as `order_hour_sin` and `order_hour_cos` have weak correlations with reorder metrics.

Table 13: Correlation Matrix

Feature	aisle	department	product	reorder_rate	add_to_cart
aisle_target_enc	1.00	0.65	0.68	0.74	-0.15
department_target_enc	0.65	1.00	0.72	0.68	-0.12
product_name_target_enc	0.68	0.72	1.00	0.80	-0.14
product_reorder_rate	0.74	0.68	0.80	1.00	-0.22
avg_add_to_cart_order	-0.15	-0.12	-0.14	-0.22	1.00

4 Confusion Matrix

Classification Metrics:

- **Accuracy:** 0.77 - Indicates 77% of predictions were correct.
- **Precision:** 0.82 - Out of all predicted positives, 82% were actual positives.
- **Recall:** 0.78 - Out of all actual positives, 78% were correctly identified as positives.
- **F1 Score:** 0.80 - The harmonic mean of precision and recall, balancing the two.

Table 14: Confusion Matrix

	Predicted: 0	**Predicted: 1**	**Total**
Actual: 0 (TN)	606,460	193,225	799,685
Actual: 1 (TP)	254,858	891,527	1,146,385
Total	861,318	1,084,752	1,946,070

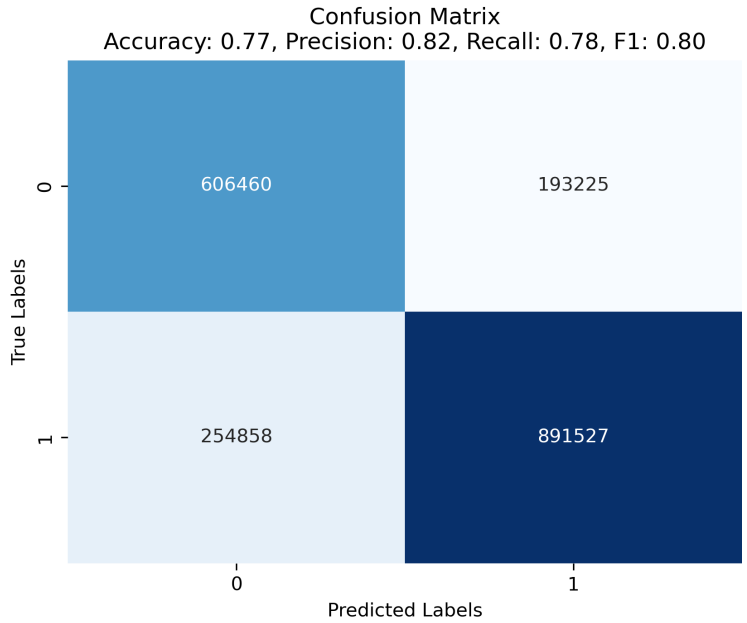

Figure 1: Confusion Matrix Visualization for Logistic Regression

5 Classification Report

Details of Performance Metrics:

- True Negatives (TN): The model correctly identified 606,460 instances as class 0.
- False Positives (FP): 193,225 instances were incorrectly predicted as class 1.
- False Negatives (FN): 254,858 instances were missed and classified as class 0.
- True Positives (TP): 891,527 instances were correctly predicted as class 1.

Table 15: Classification Report

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Class 0	0.76	0.88	0.81	799,685
Class 1	0.82	0.72	0.77	1,146,385
Avg/Total	0.79	0.77	0.79	1,946,070

6 Conclusion

The correlation matrix highlights the strongest predictors for reorder behavior, while the confusion matrix and classification metrics demonstrate the model's overall performance. The classification report confirms the balance between precision and recall, with an F1 score of 0.80 indicating robust predictions.