Explanation of Your Engineered Features

You have successfully created a wide variety of powerful features by combining the raw data with insights from your exploratory data analysis. Here is a detailed breakdown of each new feature group.

1. Global Offer Features (The "Offer DNA")

These features describe the inherent characteristics and historical performance of each offer, based on the add_event and offer_metadata files. They are "global" because they are not specific to any single customer.

- **offer_total_impressions**: The total number of times an offer has been shown to any customer in the historical log.
  - **Why it's important:** Acts as a measure of an offer's **reach or exposure**.
- **offer_total_clicks**: The total number of times an offer has been clicked by any customer in the historical log.
- **offer_historical_ctr**: The historical Click-Through Rate (total_clicks / total_impressions) for an offer.
  - **Why it's important:** This is one of your most powerful features. It's a direct measure of an offer's **popularity or appeal**.
- **offer_redemption_freq**: Indicates if an offer is single-use (1) or multi-use (2).
  - **Why it's important:** Your EDA showed that multi-use offers are significantly more likely to be clicked.
- **offer_discount_rate**: The percentage discount offered.
- **offer_type_code**: A high-level category for the offer, likely distinguishing between cashback (1) and points/miles (2).
- **offer_duration_days**: The number of days an offer is valid.

2. Global Industry Features (The "Industry DNA")

These features describe the general spending patterns within each merchant industry, based on the add_trans file.

- **industry_avg_spend**: The average transaction amount for a given industry.
- **industry_total_transactions**: The total number of transactions recorded for an industry.
- **industry_unique_products**: The number of different product IDs sold within an industry.
  - **Why they're important:** These features provide crucial context. Your model can learn, for example, that an offer in a "high average spend" industry might be more appealing.

3. Time-Based & Interaction Features

These features are created directly from the id4 (impression_timestamp) column in your main train/test sets. They capture the specific context of the interaction.

- **hour_of_day**, **day_of_week**, **is_weekend**: Basic time features that capture daily and weekly patterns.
- **hour_sin**, **hour_cos**, **day_sin**, **day_cos**: Cyclical features that help the model understand the cyclical nature of time (e.g., that hour 23 is close to hour 0).
- **offer_type_x_weekend**: An interaction feature that combines the offer type with whether it's a weekend.
  - **Why it's important:** It allows the model to learn if certain offer types are more effective on weekends.
- **offer_ctr_vs_type_avg**: Compares an offer's historical CTR to the average CTR for its offer_type_code.
  - **Why it's important:** This is a powerful feature that tells the model if an offer is an **overperformer or underperformer** compared to its peers.

4. Customer Cluster Features

These are the most advanced features you've created. They are new categorical features that group customers into distinct segments based on their underlying behaviors.

- **spending_cluster**: Groups customers based on their spending habits (e.g., "High-Value Spenders").
- **history_cluster**: Groups customers based on their past click/impression history.
- **profile_cluster**: Groups customers based on their loyalty and profile attributes.
- **engagement_cluster**: Groups customers based on how they interact with the Amex website/app.
- **holistic_cluster**: Groups customers based on all their anonymized f features combined.

5. Entity Cluster Features

These features are similar to the customer clusters, but they group the **offers** and **industries** themselves.

- **offer_cluster**: Groups offers into segments based on their combined historical performance and static details (e.g., "High CTR, Weekend-Heavy Offers").
- **industry_cluster**: Groups industries based on their collective transaction patterns (e.g., "High Volume, Low Value Industries").

By combining all these different types of features, you have created a rich, multi-dimensional dataset that gives your final model the best possible chance of finding the complex patterns needed to achieve a high score.