# 3.Complex Open-Ended Problem-Deriving Demographics(Age&Gender Prediction)

**Objective**

The aim is to **infer customer age groups (<25, 25–40, 40+) and gender (Male/Female)** using only transactional data. Since direct demographic details are unavailable, the framework leverages behavioral and purchase patterns such as product categories, spending habits, purchase frequency, and time-of-day signals.

**1. Approach and Key Signals**

Demographics are inferred through behavioral indicators derived from transactions:

* **Product Category**: Certain categories strongly map to age and gender.
* **Price Sensitivity**: High-value purchases often correlate with older, premium segments.
* **Purchase Frequency**: Younger customers tend to shop more frequently in smaller amounts.
* **Time of Purchase**: Late-night activity is a proxy for younger demographics.
* **Brand Affinity**: Electronics brands skew male, while fashion and beauty categories skew female.

**2. Age Group Framework**

Customers are segmented into three age brackets using rule-based signals:

**<25 years**

* Frequent, small-value purchases
* High activity between **10 PM–2 AM**
* Spend on gadgets, gaming, budget fashion

**25–40 years**

* Working professionals and parents
* Higher average order value
* Purchases in **Kids, Lifestyle, Home** categories
* Preference for premium branded items

**40+ years**

* Fewer but larger transactions
* Focus on **healthcare, appliances, jewelry**
* Shopping mostly during daytime

**3. Gender Framework**

Gender prediction relies on product categories and transaction behaviors:

**Male**

* Electronics, gaming, mobiles dominate
* Preference for premium brands (Apple, Sony, OnePlus)
* Higher-value, less frequent transactions

**Female**

* Beauty, fashion, footwear dominate
* Frequent mid-value purchases
* Active participation during **festival/discount events**

**4. Scoring Rules (Illustrative)**

A scoring engine assigns points to each demographic bucket. The category with the **highest cumulative score** determines the inferred demographic.

| **Rule** | **Inference** | **Score** |
| --- | --- | --- |
| >60% spend in Beauty/Apparel | Female | +2 |
| >50% spend in Kids category | Age 25–40 | +2 |
| Avg. order value > ₹80,000 | Male, 25–40 | +2 |
| Purchases between 10 PM–2 AM | Age <25 | +1 |
| >70% spend in Healthcare/Appliances | Age 40+ | +2 |
| Premium brands dominate | Male | +1 |
| >10 transactions in 2 months | Age <25 | +1 |

**5. Example Outcomes**

* **Apple/Sony products, ₹90,000+ spend, festival-driven shopping → Male, 25–40**
* **Frequent budget fashion, late-night purchases → Female, <25**
* **Healthcare + appliances, rare but ₹70,000+ spend → Male/Female, 40+**

**6. Validation Strategy**

Since explicit labels are unavailable, validation can be done once demographic data is collected:

* **Precision & Recall** for accuracy
* **Confusion Matrix** to refine rules
* **Lift vs Random Baseline** for performance gain
* **A/B Testing** with targeted campaigns

**7. Key Assumptions**

* Product-category preferences reflect demographic trends.
* High-value purchases → likely 25–40 working professionals.
* Night-time shopping → proxy for <25 segment.
* Gender-category preferences remain broadly consistent.

**8. Strategic Value**

This framework helps businesses to:

* Build demographic segments without explicit data.
* Personalize marketing (e.g., gadgets for males, beauty products for females).
* Launch age-specific offers (student discounts, family bundles).
* Optimize merchandising and inventory planning.

By combining **purchase behavior, product affinity, and spending signals**, businesses can approximate customer demographics effectively, improving personalization and engagement.

