**Methodology for Inferring User Demographics from Transactional Data (Buyhatke- Growth: Section 3)**

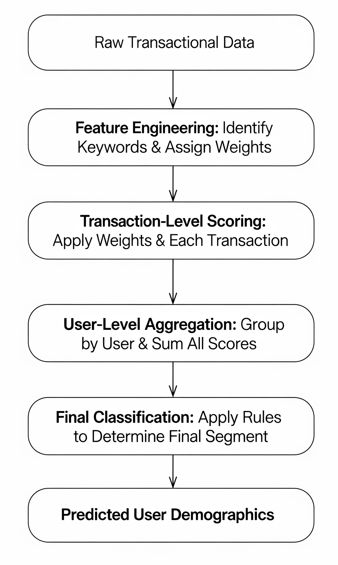
**1. Introduction & Problem Statement**

The objective is to develop a framework for inferring user demographics—specifically age group and gender—based solely on their transactional history, as explicit demographic data is often unavailable.

This document outlines the methodology for a **Python-based, rule-scoring model** that leverages signals from an audio transactional dataset. The framework's strength lies in the **synergy** of combining multiple data points to classify users into predefined buckets:

* **Gender:** Male, Female
* **Age Groups:** <25, 25–40, 40+

The initial prototype focuses on **product signals** (derived from keywords in product names), but the methodology is designed to incorporate **behavioural signals** (such as spending power and purchase frequency) for greater accuracy. This document details the step-by-step reasoning, weighting assumptions, formulas, and the proposed validation approach for the model.



Flowchart

**2. Core Assumptions**

This model is built upon a set of foundational assumptions about consumer behaviour:

1. **Product Preference Correlation:** It is assumed that preferences for certain product categories, brands, and features (like colour) correlate strongly with specific demographic profiles.
2. **Keyword-as-Proxy:** Keywords found within product names (e.g., "gaming," "portable," "amplifier") are effective proxies for identifying the product's intended audience.
3. **Aggregated Behaviour:** A user's demographic profile is not determined by a single purchase but is better represented by the aggregation of all their transactional signals over time.

**3. Step-by-Step Methodology and Reasoning**

The process of classifying a user involves four key steps, moving from individual transaction analysis to a final user-level demographic prediction.

**Step 1: Signal Identification and Weighting**

The first step is to define the signals that will be extracted from the data. We identified keywords in product names that are likely to indicate a specific demographic. Each keyword is assigned a weight, with positive values leaning towards one group and negative values towards another.

**A) Gender Signals:** For gender, a positive score indicates a higher likelihood of being Male, while a negative score indicates a higher likelihood of being Female.

| **Keyword** | **Weight** | **Rationale** |
| --- | --- | --- |
| gaming, headset | +1.5 | The gaming accessories market has historically skewed towards a male demographic. |
| pink, rose gold | -2.0 | Color preference can be a strong, though not definitive, signal for gender. |

**B) Age Group Signals:** For age, each keyword contributes a score to one or more of the three age buckets.

| **Keyword** | **<25 Score** | **25–40 Score** | **40+ Score** | **Rationale** |
| --- | --- | --- | --- | --- |
| gaming | +2.0 | 0 | 0 | Strong indicator for a younger demographic involved in interactive entertainment. |
| portable | 0 | +1.0 | 0 | Portable audio is popular across a wide range, but peaks with mobile, commuting adults. |
| amplifier, audiophile | 0 | 0 | +2.5 | High-cost, specialized audio equipment is often a hobby for an older demographic with more disposable income. |

**Step 2: Transaction-Level Scoring**

Each transaction is scored individually based on the keywords found in its product\_name.

**Formulas:** For a single transaction, the scores are calculated as follows:

* *Gender Score* = Σ (Weight of each identified gender keyword)
* *Age <25 Score* = Σ (Score of each identified '<25' keyword)
* *Age 25-40 Score* = Σ (Score of each identified '25-40' keyword)
* *Age 40+ Score* = Σ (Score of each identified '40+' keyword)

**Step 3: User-Level Score Aggregation**

To build a complete profile, the scores from all transactions made by a single user are aggregated. This is achieved by grouping the dataset by the user identifier and summing the scores from Step 2.

* *Total User Male Score* = Σ (*Gender Scores* from all user transactions)
* *Total User Age <25 Score* = Σ (*Age <25 Scores* from all user transactions)
* ...and so on for the other age groups.

This ensures that a user's profile is built from their entire purchase history, making the prediction more robust.

**Step 4: Final User Classification**

The final step is to assign a demographic bucket to each user based on their aggregated scores.

1. **Gender Classification:** The sign of the *Total User Male Score* determines the predicted gender.
   * If *Total User Male Score > 0*, Predicted Gender = **Male**.
   * Otherwise, Predicted Gender = **Female**.
2. **Age Group Classification:** The user is assigned to the age group with the highest total score.
   * Predicted Age Group = The group (<25, 25-40, or 40+) with the maximum score.

| **User** | **Predicted Gender** | **Predicted Age Group** |
| --- | --- | --- |
| user-1 | Female | age<25\_score |
| user-10 | Male | age<25\_score |
| user-100 | Female | age<25\_score |
| user-1000 | Female | age<25\_score |
| user-10000 | Female | age25-40\_score |

TABLE - DEMO FINAL CLASSIFICATION(in code)

**4. Validation Approach**

Since the ground truth (actual user demographics) is not available, this section describes how the model's performance *would* be validated if such data were provided.

1. **Confusion Matrix:** A confusion matrix would be the primary tool to visualize performance. It's a table that compares the model's predictions against the actual labels, showing the number of True Positives, True Negatives, False Positives, and False Negatives for each class.
2. **Precision and Recall:** These key metrics would be calculated from the confusion matrix.
   * **Precision:** Measures the accuracy of the predictions. For example, "Of all the users we predicted as 'Male', what percentage actually were male?"
   * **Recall:** Measures the model's ability to find all relevant instances. For example, "Of all the users who are actually in the '40+' age group, what percentage did our model correctly identify?

3. **Lift Over Baseline:** The model's performance would be compared to a random baseline. For gender, a random guess would be 50% accurate. A successful model must demonstrate a significant "lift," or improvement, over this baseline.