

# **Sample Case Study Report – Data Driven Modelling**

**Course:** Data Driven Modelling

**Assessment Type:** CIA-1

**Title:** Design and Documentation of a Primary Dataset for Predicting Product Popularity on Amazon India

**Student Details:** Faustena S (2548313)

## **1. Introduction and Problem Definition**

The Indian e-commerce beauty and personal care market is growing rapidly. Many small sellers and emerging brands list makeup, skincare, and cosmetic products on Amazon India. However, choosing the right price and discount level is difficult — wrong decisions lead to low sales and wasted inventory.

Products that receive a large number of customer ratings are usually the ones that sell well and gain trust quickly. The objective of this case study is to design and document a primary dataset collected from Amazon India that can support data-driven modelling to predict whether a beauty product is likely to become **popular** (high sales potential).

This problem is inductive in nature — patterns of popularity are expected to emerge from observed pricing, discount, and rating data rather than from fixed rules decided in advance.

## **2. Data Source and Data Collection Plan**

The dataset is primary data collected directly from the Amazon India website.

**Source:** Amazon.in search results page for beauty, makeup, and skincare products (publicly visible page, no login required).

**Collection method:** Manual extraction by the student on 15 February 2026 using a browser in private mode.

**Target items:** Beauty and makeup related products visible in a broad search.

**Sampling strategy:** Complete capture of visible products in one search session (no random sampling, no filtering).

**Duration:** Single snapshot collected in one continuous session (approximately 45–60 minutes).

### **3. Rules and Steps for Data Collection**

#### **Data Collection Rules:**

1. Every product shown in the search results must be recorded (no skipping).
2. All values must be copied exactly as displayed (price, original price, discount, rating, number of ratings).
3. No values should be guessed or corrected during collection.
4. Product names should be recorded fully to allow duplicate detection later.

#### **Steps Followed:**

1. Open Amazon.in in private browsing mode.
2. Perform a broad search for beauty/makeup products.
3. Scroll through the results page and copy each product's information row by row into Excel.
4. After collection, randomly select 8–10 products and click into their detail pages to verify price and rating match.
5. Save the file immediately as amazondmm1.xlsx.

### **4. Dataset Description**

The raw dataset consists of the following attributes:

- Product Name
- Price (INR)
- Original Price (INR)
- Discount Percentage
- Rating (Max 5)
- Number of Ratings

**Dataset shape (raw):** 155 rows × 6 columns

**Label distribution (before any cleaning):**

- Not Popular (0): 107 (69.0%)
- Popular (1): 48 (31.0%)

## 5. Data Cleaning and Preprocessing Strategy

During initial inspection, the following data quality issues were observed:

- Duplicate products (same name and nearly identical details)
- Missing values in Original Price and Discount Percentage columns
- Commas in large numbers in the Number of Ratings column (e.g. 40,623)

### Cleaning Actions:

- Removed duplicate rows based on Product Name (kept first occurrence).
- For missing Original Price → set Discount Percentage = 0.
- Removed commas from Number of Ratings and converted to numeric type.
- Created two derived features:
  - $\text{Price\_to\_Original\_Ratio} = \text{Price} / \text{Original Price}$
  - $\text{Savings\_INR} = \text{Original Price} - \text{Price}$
- No rows were removed due to missing values in the target variable (Number of Ratings was present in all rows).

After cleaning, the working dataset retained 155 rows (no row-level deletion was needed for modelling).

## 6. Labeling Strategy

The dataset is labelled to support binary classification.

**Label Objective:** Classify each product as Popular or Not Popular.

### Label Definition:

- Popular (1): Number of Ratings  $\geq 5,000$
- Not Popular (0): Number of Ratings  $< 5,000$

This rule-based labeling is transparent and directly reflects real customer engagement behavior on Amazon. Products with 5,000+ ratings are generally proven best-sellers in the Indian beauty category.

## 7. Benchmarking and Evaluation Plan

**Objective:** Evaluate how well pricing, discount, and rating features can predict product popularity.

**Features used:** Price (INR), Original Price (INR), Discount Percentage, Rating (Max 5), Price\_to\_Original\_Ratio, Savings\_INR

**Train / Test split:** 116 training rows | 39 test rows (approx. 75/25)

**Baseline benchmark:** Majority class classifier (predict Not Popular for all products)

**Models evaluated:**

- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- K-Nearest Neighbours
- SVM

**Evaluation metrics:** Accuracy, ROC-AUC, Precision, Recall, F1-score

**Key results (from pipeline):** Best model: **Random Forest**

- Test Accuracy: 84.6%
- ROC-AUC: 0.894

**Classification report (Random Forest on test set):**

	precision	recall	f1-score	support
Not Popular (0)	0.92	0.85	0.88	27
Popular (1)	0.71	0.83	0.77	12
accuracy			0.85	39
macro avg	0.82	0.84	0.83	39
weighted avg	0.86	0.85	0.85	39
CROSS-VALIDATION SUMMARY TABLE				
Model	CV Acc	Test Acc	ROC-AUC	
Logistic Regression	0.682 ± 0.028	0.667	0.722	
Decision Tree	0.707 ± 0.083	0.769	0.719	
Random Forest	0.690 ± 0.100	0.846	0.894	
Gradient Boosting	0.681 ± 0.093	0.795	0.836	
K-Nearest Neighbours	0.697 ± 0.080	0.795	0.781	
SVM	0.724 ± 0.043	0.769	0.802	

## **8. Intended Use of the Dataset**

The dataset can be used for:

- Helping small sellers test price and discount combinations before listing large quantities.
- Supporting new beauty brands in deciding competitive launch prices on Amazon India.
- Performing exploratory analysis of pricing patterns in the Indian beauty e-commerce segment.
- Serving as a learning resource for students studying e-commerce analytics and classification problems.

## **9. Conclusion**

This case study presents a complete end-to-end process for designing and documenting a primary dataset for data-driven modelling. Emphasis is placed on realistic data collection from a public e-commerce platform, systematic cleaning, clear rule-based labeling, and fair model benchmarking. The dataset and pipeline provide a solid foundation for further analysis of product popularity in the fast-growing Indian beauty market.

