# **Sprint 2 Documentation — Price-First Smart**

# **Substitution (Cart Optimization)**

**Project:** DiscountMate – Smart Substitution  
**Sprint:** 2 (Price-First Model + Model EDA)  
**Author:** Sharon Roy (ML)  
**Contributors:** Bailey (data consolidation), Bethmi (similarity inputs), Raj (evaluation guidance)  
**Repo Notebook:** Sprint\_2\_CAP.ipynb  
**Data (example):** smart\_substitution\_dataset 8.csv

## **1. Executive Summary**

This sprint delivers a **price-first substitution pipeline** that recommends a **cheaper, comparable** product when a shopper selects an item. We implemented:

* A **deterministic rules baseline** that enforces comparability (subcategory, unit, size) and picks the **cheapest** valid option.
* A lightweight **price-aware ranker** (Logistic Regression) to re-order comparable candidates using **ΔPrice**, **ΔPPU**, text similarity, and size/brand signals.
* A **model-only EDA** suite to report **coverage, % cheaper, savings per item, acceptance rate**, and subcategory diagnostics.

**Headline (latest run, n≈468):** Coverage **94%**, **37%** of items had a cheaper alternative, median ΔPrice **$0.00**, mean ΔPrice **−$1.16**, acceptance **49%**.  
Next sprint focuses on **cheaper-only gating**, **larger candidate pool**, and **subcategory-specific tolerances** to lift savings and acceptance.

## **2. Objective**

Optimise cart spend by suggesting a **like-for-like, lower-cost** alternative that preserves usability:

* Same **subcategory** and **unit type** (if present),
* **Comparable size** (same size band or high size proximity),
* **Transparent savings** (ΔPrice and ΔPPU displayed).

## **3. Data Overview**

* **Source:** Team-merged Sprint dataset (Bailey).
* **Key fields (auto-mapped in notebook):**  
  name, brand, subcategory, unit\_type, std\_item\_size, size\_band (derived if missing), pack\_count, price *(if present)*, price\_per\_unit (PPU).
* **Price handling:** Prefer shelf **price**. If missing, estimate **pack price** = PPU × std\_item\_size × pack\_count (clearly labeled as estimate).

## **4. Method & Architecture**

## **4.1 Rules Baseline (Price-First)**

1. **Comparable set:**
   * same subcategory,
   * same unit\_type (if present),
   * **size compatibility:** same size\_band ±1, or **size\_prox ≥ 0.80** when sizes exist.
2. **Selection:** choose the **cheapest** candidate by **item price** (fallback to **PPU** if price missing).
3. **Explainability:** tags like Same unit, Same size band, Cheaper by X%, plus **ΔPrice/ΔPPU**.

## **4.2 Model Layer (Price-Aware Ranker)**

* **Features:** ΔPrice, ΔPrice%, ΔPPU, ΔPPU%, TF-IDF **text similarity**, size\_prox, same\_band, brand\_eq.
* **Weak labels:** acceptance rule (same subcat/unit, size/band compatible, PPU within ±20%, weak text gate).
* **Model:** LogisticRegression (class\_weight="balanced"), **P(good)** for re-ranking; ties broken by **greater monetary saving**.

## **4.3 Acceptance Rule (for evaluation only)**

A swap is **accepted** if: same subcategory, unit matches (if present), **same band** or **size\_prox ≥ 0.80**, and **PPU** within **±20%** of the query (very weak text only when brand aligns).

## **5. How to Run (Notebook Flow)**

1. **Part A — Load & Prepare:** column mapping, derive size band, compute estimated price if needed.
2. **Part B — Price-First Baseline:** cheapest\_substitute(qidx, prefer="item") returns one valid, cheapest substitute with **ΔPrice/ΔPPU**.
3. **Part C — Model Build:** train price-aware ranker; rank\_with\_model(qidx, k=3) returns top-k with savings.
4. **Model-Only EDA (M-EDA0..5):** coverage, % cheaper, median/mean ΔPrice, acceptance, by-subcategory breakdown, failure buckets, savings quantiles.

## **6. Results (Model-Only EDA — latest run)**

* **Sample:** 468 queries
* **Coverage:** **94%** (model returns a top-1 in most cases)
* **Cheaper alternative found:** **37%** of items
* **Savings per item (ΔPrice):** median **$0.00**, mean **−$1.16** *(positive = saving)*
* **Unit savings (ΔPPU):** median **+0.0004**
* **Acceptance rate:** **49%**

**Savings distribution (ΔPrice):**  
P10 = **−$8.08**, P25 = **−$1.30**, **Median = $0.00**, P75 = **+$1.60**, P90 = **+$5.64**.

**Strong categories:** cheese blocks, diffusers, dishwashing  
**Weak categories:** baby wipes, bakery snacks, bleach & stain removers *(small n per subcat; directional)*

## **7. Interpretation**

* The system is **reliable and explainable** (good coverage, clear tags).
* **Savings are uneven**: many price-neutral items (median $0.00), and a tail of **more-expensive picks** dragging the mean below zero.
* Acceptance shows **half** of model picks meet strict comparability + PPU gating.

## **8. Recommendations (Next Sprint)**

1. **Cheaper-only gate at inference:** prefer cand\_price < query\_price; only fall back when no cheaper comparable exists.
2. **Broaden candidate pool:** raise k\_cand (e.g., **60 → 120**) to surface more bargains.
3. **Per-subcategory tuning:**
   * *Commodities/snacks:* tighten PPU window; keep strict size band.
   * *Premium/beauty/spirits:* widen PPU window; allow size\_prox ≥ 0.70 with band guard.
4. **Price QA:** ensure shelf price coverage; when estimating from PPU×size×pack, label clearly and check units.
5. **Light text normalisation:** two-token brands, remove generic tokens (pack, ml, classic, etc.) to reduce off-topic candidates.
6. **Optional model upgrade:** switch to **LambdaMART (LightGBM)** optimised for **NDCG@3** to improve top-k ordering.

## **9. Key Parameters (Reproducibility)**

* **Size bands:** Small <250, Medium 250–749, Large ≥750 (derived if missing)
* **Size proximity:** size\_prox\_min = 0.80
* **PPU tolerance (acceptance):** ±20%
* **Text features:** TF-IDF min\_df=2, n-grams (1,2)
* **Ranker:** LogisticRegression (max\_iter=500, class\_weight="balanced")
* **Candidate pool:** k\_cand = 60 *(suggest 120)*
* **Random seed:** 42

## **10. File Map (Sprint 2)**

* **Notebook:** Sprint\_2\_CAP.ipynb
* **Data (example):** smart\_substitution\_dataset 8.csv

## **11. Limitations**

* **Price sparsity:** if shelf price is missing, ΔPrice uses an estimate.
* **Heterogeneous subcategories:** wide pack/size variation reduces like-for-like availability.
* **Weak labels:** acceptance is a proxy; a small human-labelled set would improve training.

## **12. Contacts / Credits**

* **Sharon Roy** — Sprint 2 implementation & coordination
* **Bailey** — dataset consolidation
* **Bethmi** — similarity inputs
* **Raj** — brand & evaluation guidance