

Analiză preferințe produse și sisteme de recomandare

Antoniev Valeriu-Gabriel și Roșu Ioan

January 10, 2026

Abstract

Acest raport prezintă dezvoltarea și calitatea unor modele de învățare automată (Regresie logistică și Naive Bayes) aplicate pe un set de date de bonuri fiscale. Obiectivele au fost: predicția achiziției unui produs specific, Crazy Sauce, considerând doar bonurile care conțin Crazy Schnitzel, recomandarea de sosuri și generarea unui ranking de produse pentru upsell. Rezultatele experimentale demonstrează o acuratețe superioară a metodelor ML (99.7%) comparativ cu baseline-urile statistice (74%).

1 Descrierea Problemei și a Dataset-ului

1.1 Context

Scopul este de a înțelege comportamentul consumatorului prin analiza asocierilor dintre produse. Dataset-ul provine dintr-un mediu de retail (fast-food).

1.2 Preprocesarea datelor

Datele inițiale au constat într-o listă de produse, unde fiecare rând reprezenta un item scanat. Pașii de preprocesare efectuați au fost:

1. **Agregarea:** Gruparea datelor după `id_bon` pentru a reconstrui coșurile de cumpărături.
2. **Transformare:** Rescrierea listei de produse într-o matrice de frecvență, unde coloanele sunt produsele și rândurile sunt bonurile.
3. **Adăugare de attribute:**
 - Obținerea atributelor `day_of_week` și `is_weekend`.
 - Calculul `cart_size` și `total_value` pentru fiecare bon.
4. **Filtrare:** Pentru prima sarcină, s-a păstrat un subset de date condiționat de prezența produsului "Crazy Schnitzel".
5. **Normalizare:** Standardizarea datelor ($x' = \frac{x-\mu}{\sigma}$) pentru a asigura convergența algoritmului Gradient Descent.

2 Metodologie

2.1 Regresia Logistică (Task 1 și 2)

Am implementat regresia logistică în Python folosind numpy. Putem analiza ponderile w pentru a înțelege ce produse influențează decizia.

Configurare Experimentală:

- **Algoritm:** Gradient ascendent.
- **Funcția de Cost (Log-Likelihood):**
$$\ell(w) = \sum_{i=1}^n (y^{(i)} \ln \sigma(w \cdot x^{(i)}) + (1 - y^{(i)}) \ln(1 - \sigma(w \cdot x^{(i)})))$$
- **Learning Rate:** 1.0.

2.2 Naive Bayes / Ranking Probabilistic (Task 3)

Pentru ranking-ul produselor modelul estimează $P(\text{Produs}|\text{Context})$ folosind frecvențele produselor din date. S-a utilizat regula lui Laplace pentru a evita problema apariției de probabilități nule.

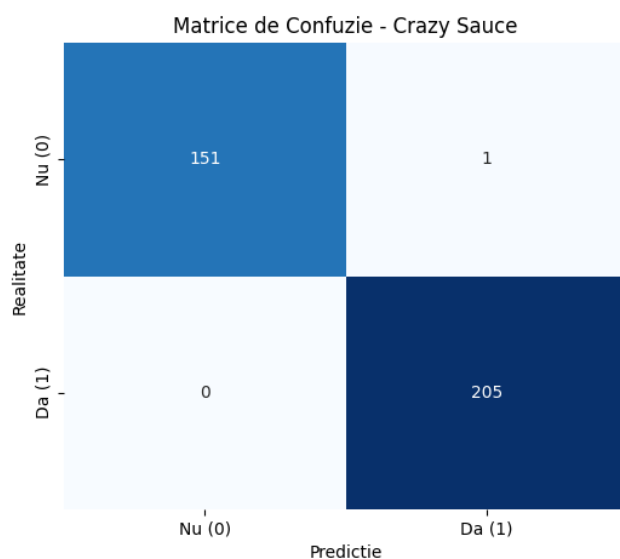
3 Rezultate Experimentale

3.1 Task 1: Predicția Crazy Sauce

Modelul a fost antrenat pe 1426 bonuri și testat pe 357. **Performanță:**

- **Acuratețe:** 99.72%
- **Matrice de Confuzie:** TP=205, TN=151, FP=1, FN=0.

Modelul separă aproape perfect clasele. O singură eroare (False Positive) a fost înregistrată.



Interpretarea Ponderilor: Graficul de mai jos ilustrează cei mai importanți factori. Se observă că `cart_size` are o influență pozitivă masivă (coșurile mari tind să aibă sos), în timp ce prezența altor sosuri (*Cheddar*, *Garlic*) scade drastic probabilitatea (efect de substituție).

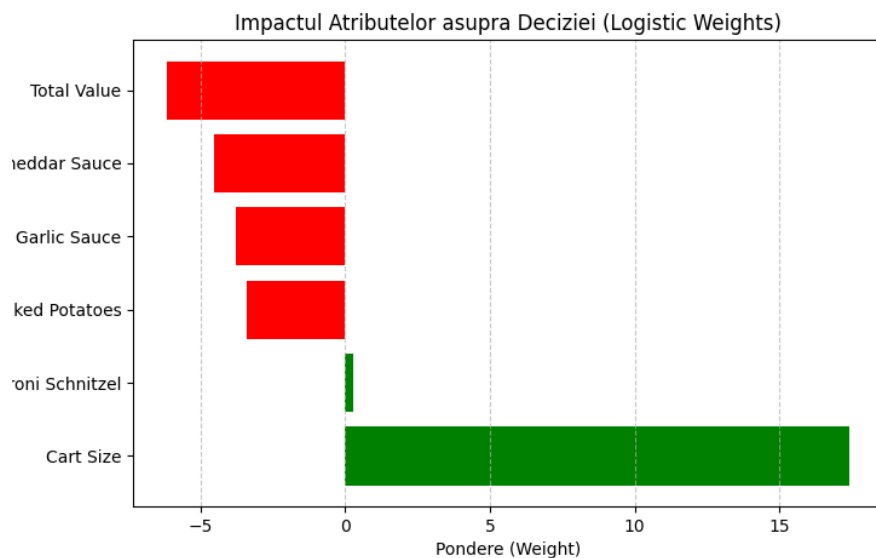


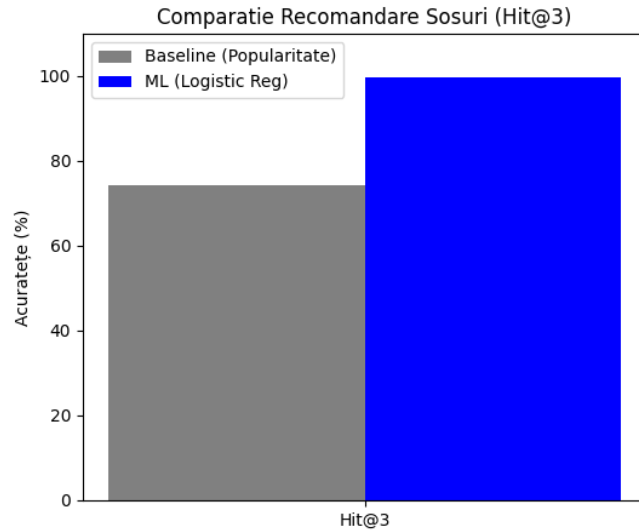
Figure 1: Ponderile Regresiei Logistice. Valorile negative indică produse care se exclud reciproc cu Crazy Sauce.

3.2 Task 2: Recomandare Sosuri (Multi-label)

S-a antrenat câte un model independent pentru fiecare tip de sos. Evaluarea s-a făcut folosind metrica *Hit@3* (dacă sosul real se află în top 3 recomandări).

Metoda	Hit@3 Accuracy
Baseline (Popularitate Globală)	74.19%
Regresie Logistică (Contextual)	99.76%

Table 1: Comparatie Baseline vs ML



3.3 Task 3: Ranking și Upselling

Folosind metoda probabilistică (tip Naive Bayes) și evaluarea *Leave-One-Out* pe 1259 bonuri, am obținut următoarele rezultate:

- **Hit@1:** 26.77% (Produsul exact a fost ghicit în 1 din 4 cazuri).
- **Hit@3:** 50.75% (Produsul exact a fost ghicit în aproximativ jumătate din cazuri).
- **Hit@5:** 61.16% (Produsul s-a aflat în top 5 sugestii în majoritatea cazurilor).

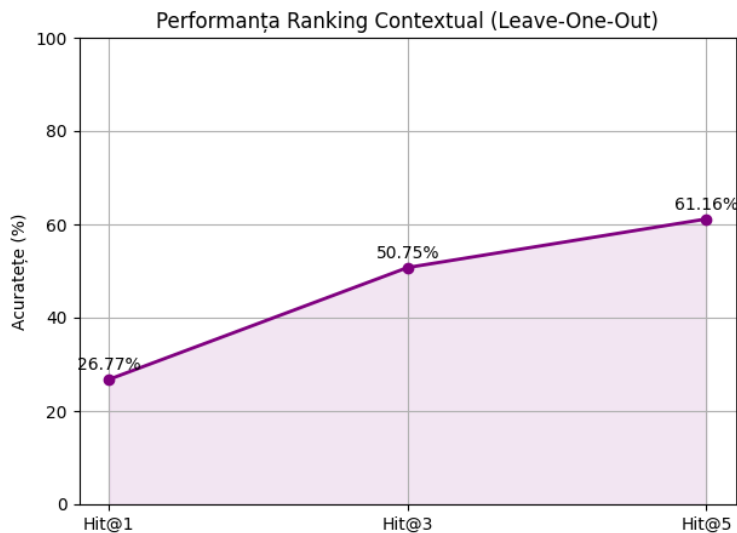


Figure 2: Curba de performanță Hit@K. Acuratețea crește semnificativ pe măsură ce creștem fereastra de recomandare K.

4 Concluzii și Direcții Viitoare

Concluzii:

1. Regresia Logistică este extrem de eficientă pentru acest dataset, probabil datorită corelațiilor puternice și structurate dintre produsele de tip "Meniu" (ex: Schnitzel implică aproape întotdeauna un sos și o băutură).
2. Efectul de substituție (Clientul cumpără Garlic Sauce → Probabilitatea de Crazy Sauce scade) a fost captat corect de ponderile negative.
3. Recomandarea contextuală este net superioară simplei popularități, crescând rata de succes de la 74% la 99% în cazul sosurilor.

Direcții de Îmbunătățire:

- **Date Temporale:** Analiza secvențială (dacă am avea ID de client recurent) pentru a prezice ce va cumpăra data viitoare.
- **Regularizare L1 sau L2.**