# Practical Assignment – Machine Learning

Mazilu Cosmin-Alexandru
Mardare Andrei Daniel

January 16, 2026

## 1 Problemă & Dataset

Datele provin din bonuri fiscale; fiecare linie reprezintă un produs dintr-un bon. Interval: 2025-09-05 – 2025-12-03, ~7,869 bonuri, 28,039 linii, 59 produse unice. Coș mediu: 3.56 produse/bon (max 26). Lista sosuri standalone: Crazy, Cheddar, Extra Cheddar, Garlic, Tomato, Blueberry, Spicy, Pink.

## 2 Preprocesare

Conversie `data_bon` la `datetime`; derivare `day_of_week` (1–7), `hour_of_day`, `is_weekend`.

Agregări pe bon: `cart_size`, `distinct_products`, `total_value`.

Vector produse: număr de apariții per produs (`binary_counts` opțional). Pentru modele de sos, coloana sosului curent se elimină (evităm scurgeri).

Interacțiuni opționale: (Crazy Schnitzel, French fries/Baked potatoes/Aqua Carpatica).

Split pe bonuri: bonurile au fost despărțite in ordine temporala: primele 80% din bonuri au fost folosite pentru antrenare si restul de 20% au fost folosite pentru testare.

Ranking: se construiesc coșuri parțiale eliminând aleator un produs eligibil; prețul mediu per produs folosit în scor.

## 3 Modele

**Regresie Logistică (custom GD)** Implementare proprie cu standardizare internă, gradient descent batch, oprire la toleranță, L2 opțional. Folosește doar `numpy`.

**Regresie Logistică (sklearn)** Pipeline `StandardScaler` + `LogisticRegression` pentru referință.

**Per-sauce (2.2)** Un model logistic per sos; metricile se calculează pe test; recomandarea Top-K ordonează sosurile după probabilități prezise, comparat cu baseline de popularitate.

**Ranking upsell** Bernoulli Naive Bayes (custom) pe prezență de produse; scorul final: $score(p \mid coș) = P(p \mid coș) \cdot price(p)$. Baseline: popularitate globală și venit global (contor × preț).

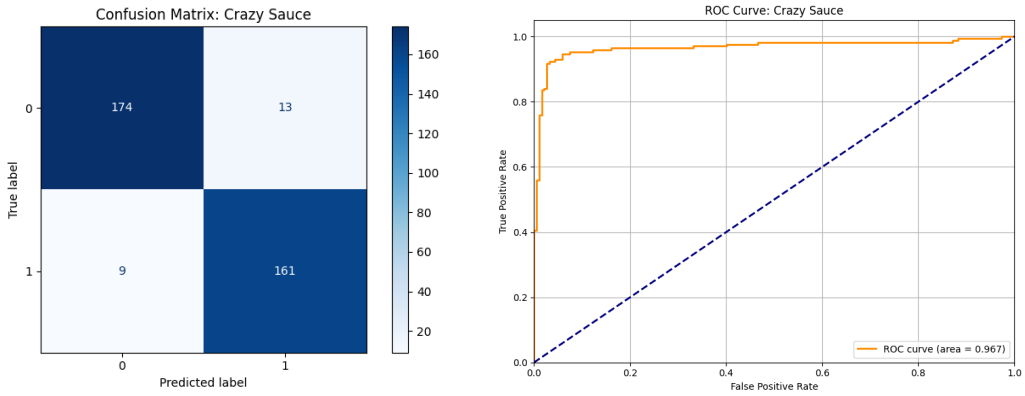


Figure 1: Matricea de confuzie și Curba ROC pentru predicția Crazy Sauce

# 4 Rezultate

## 4.1 2.1 Crazy Sauce (coș conține Crazy Schnitzel, split temporal 80/20)

| Model | Acc | Prec | Rec | F1 | ROC-AUC |
|---|---|---|---|---|---|
| Majority (tot timpul 1) | 0.476 | 0.476 | 1.000 | 0.645 | 0.500 |
| LogReg sklearn | 0.933 | 0.906 | 0.959 | 0.931 | 0.962 |
| LogReg custom GD | **0.938** | 0.925 | 0.947 | **0.936** | **0.967** |

Matrice confuzie (custom GD): $\begin{bmatrix} 174 & 13 \\ 9 & 161 \end{bmatrix}$ (TN, FP / FN, TP).

Coeficienți majori (custom GD): pozitivi — cart_size, distinct_products, total_value, Baked potatoes, băuturi cola; negativi — alte sosuri standalone (Cheddar, Garlic, Blueberry, Tomato, Spicy, Pink) și produse deja cu sos (Crazy Fries cu Cheddar).

## 4.2 2.2 Per-sauce + recomandare (split temporal 80/20, Top-3)

| Sos | Acc | Prec | Rec | ROC-AUC |
|---|---|---|---|---|
| Crazy Sauce | 0.681 | 0.334 | 0.759 | 0.751 |
| Cheddar Sauce | 0.867 | 0.653 | 0.142 | 0.760 |
| Extra Cheddar Sauce | 0.996 | 0.000 | 0.000 | 0.956 |
| Garlic Sauce | 0.895 | 0.375 | 0.097 | 0.760 |
| Tomato Sauce | 0.974 | 1.000 | 0.024 | 0.809 |
| Blueberry Sauce | 0.673 | 0.130 | 0.409 | 0.706 |
| Spicy Sauce | 0.811 | 0.110 | 0.397 | 0.739 |
| Pink Sauce | 0.974 | 0.167 | 0.061 | 0.807 |

Recomandare Top-3 (baskets cu sos în test, $n = 829$): Hit@3 = 0.744 vs baseline popularitate 0.726; Precision@3 = 0.270 vs 0.268.

## 4.3 Ranking upsell (coș parțial, min 20 apariții produs, split temporal 80/20)

| K | Model Hit@K | Popularitate Hit@K | Venit Hit@K |
|---|---|---|---|
| 1 | 0.225 | 0.093 | 0.109 |
| 3 | 0.387 | 0.217 | 0.182 |
| 5 | 0.468 | 0.320 | 0.275 |

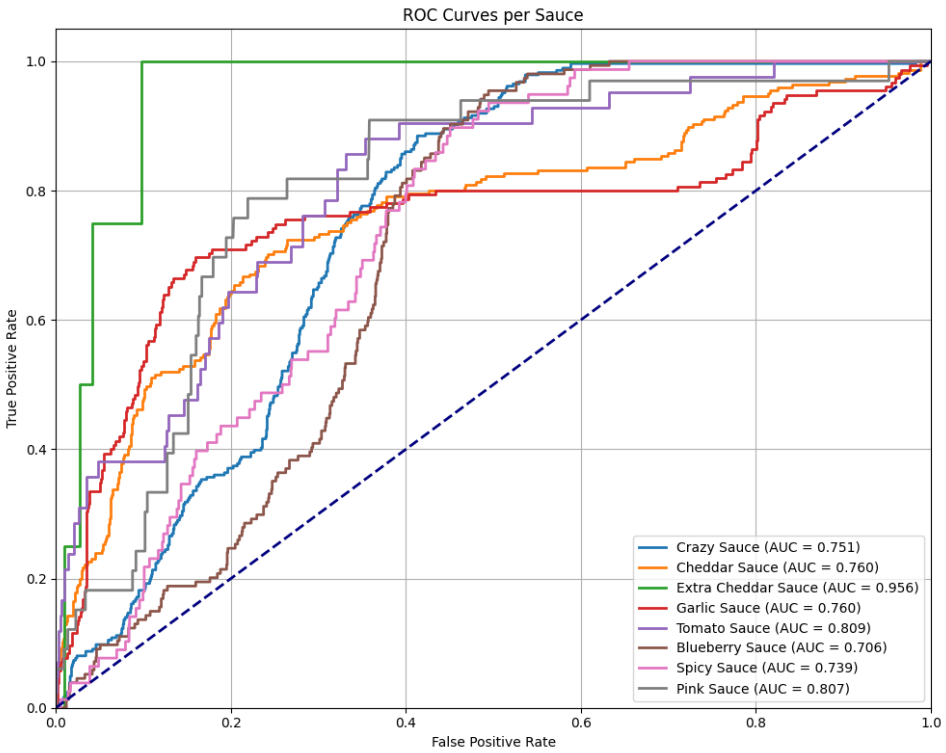


Figure 2: Curbe ROC comparative pentru modelele per-sos

Evaluat pe 1,518 coșuri parțiale. Modelul NB + preț depășește consistent bazele.

## 5 Concluzii & îmbunătățiri

Modelul logistic custom atinge performanța sklearn; principalele semnale sunt dimensiunea/varietatea coșului și substituția între sosuri.

Recomandarea per-sos depășește popularitatea, dar sosurile rare (Extra Cheddar, Pink) au recall mic; ar ajuta undersampling/oversampling sau modele ierarhice.

Ranking NB + preț bate popularitatea/venitul, dar poate fi rafinat cu modele secvențiale (item2vec), factorization machines, sau calibrare mai bună a probabilităților.

Grafică potențială: ROC-uri per sos, matrice de confuzie pentru 2.1, curbe Hit@K; pot fi generate din codul existent pentru completarea raportului.

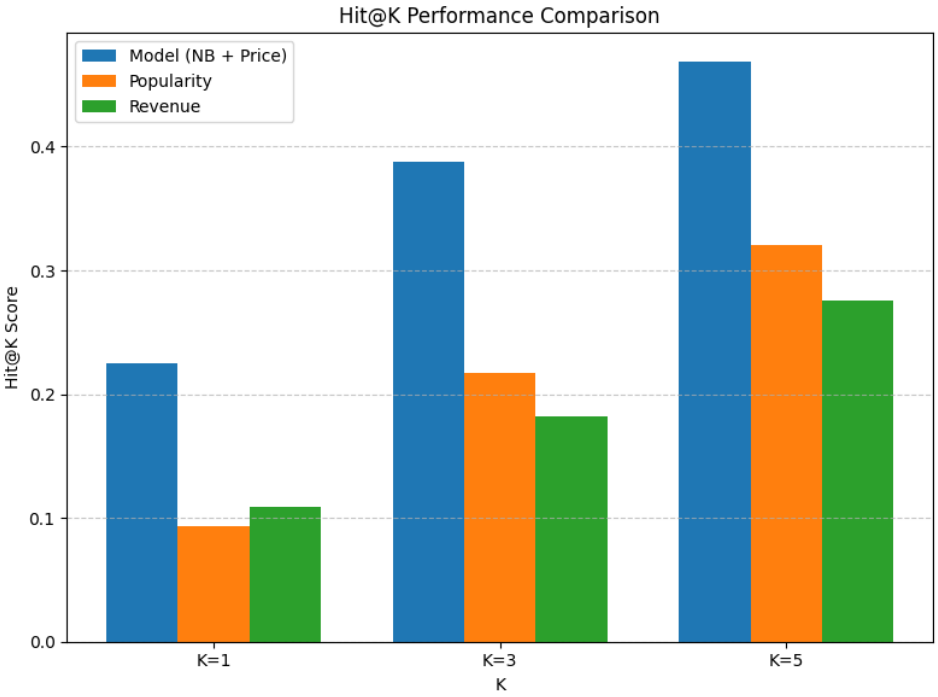


Figure 3: Performanța Hit@K pentru modelul de ranking vs baseline