HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

H&M Personalized Fashion Recommendations

GVHD: Ngô Văn Linh

Nhóm 4 :

Lương Sỹ Linh	20225029
Nguyễn Xuân Bình	20224930
Lê Bá Minh Phúc	20225065
Lương Thái Khang	20224866

ONE LOVE. ONE FUTURE.

Giới thiệu về đề tài

H&M là một tập đoàn bán lẻ thời trang lớn với hàng nghìn cửa hàng và kênh bán online, nơi khách hàng phải đối mặt với số lượng sản phẩm rất lớn. Nếu không được gợi ý phù hợp, khách khó tìm nhanh món đồ mình thích và có thể bỏ dở việc mua sắm. Vì vậy H&M tổ chức cuộc thi trên Kaggle nhằm xây dựng **hệ thống gợi ý sản phẩm cá nhân hóa**, giúp đề xuất những món đồ phù hợp cho từng khách hàng dựa trên dữ liệu lịch sử mua hàng và thông tin mô tả sản phẩm.

Chi tiết bài toán

Input:

- Lịch sử giao dịch/ mua hàng của khách (transactions)
- Thông tin khách hàng (tuổi, khu vực,…) (customers)
- Thông tin chi tiết về sản phẩm (articles)

Output :

- Với mỗi khách hàng, dự đoán ra 12 sản phẩm mà họ có khả năng mua trong tuần tiếp theo, dựa trên lịch sử và metadata kể trên. Kết quả được chấm bằng chỉ số MAP@12 (Mean Average Precision @ 12) – tức là vừa cần gợi ý đúng, vừa cần xếp hạng hợp lý những sản phẩm liên quan nhất.

Chi tiết bài toán

Hàm đánh giá:

$$MAP@12 = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{\min(m,12)}\sum_{k=1}^{\min(n,12)}P(k)\times rel(k)$$

trong đó:

- U là số lượng khách hàng,
- $P(k)$ là độ chính xác (precision) tại vị trí cắt k,
- n là số lượng dự đoán cho mỗi khách hàng,
- m là số lượng nhãn đúng (ground truth) cho mỗi khách hàng,
- $rel(k)$là một hàm chỉ thị nhận giá trị 1 nếu item tại hạng k là một nhãn liên quan/đúng, và 0 nếu không.

Chi tiết bài toán

Input: (3 file)

File	Vai trò / Ý nghĩa chính	Ghi chú quan trọng
transactions_train.csv	Lịch sử giao dịch (mua hàng) của khách	Dùng để tạo feature user–item, sequence mua sắm, tần suất, recency, giá, kênh mua
customers.csv	Thông tin thuộc tính khách hàng	Dùng để tạo feature user profile (tuổi, khu vực, độ "active", status hội viên…)
articles.csv	Thông tin chi tiết sản phẩm / mặt hàng	Dùng để tạo feature item profile (nhóm sản phẩm, màu sắc, bộ sưu tập, mô tả…)
sample_submission.csv	Mẫu định dạng file nộp kết quả (submission)	Với mỗi customer_id phải dự đoán 12 article_id cách nhau bởi dấu cách

Chi tiết bài toán

Schema transactions_train.csv

Cột	Kiểu dữ liệu	Ý nghĩa
t_dat	date	Ngày giao dịch (ngày khách mua hàng)
customer_id	string (hex)	ID khách hàng (đã mã hoá)
article_id	int	ID sản phẩm được mua
price	float	Giá bán (đã chuẩn hoá/scale theo H&M)
sales_channel_id	int	Kênh bán: 1 = online, 2 = cửa hàng (store)

Chi tiết bài toán

Schema customers.csv

Cột	Kiểu dữ liệu	Ý nghĩa
customer_id	string (hex)	ID khách hàng (khóa chính, join với transactions_train)
age	int	Tuổi (có thể có giá trị thiếu)
club_member_status	category	Trạng thái hội viên (thành viên, thường, đã rời, …)
fashion_news_frequency	category	Tần suất nhận tin thời trang (never, monthly, regular, …)
postal_code	string	Mã bưu chính / khu vực sinh sống của khách
FN	int (0/1)	Cờ (flag) liên quan đến chương trình khách hàng thân thiết / newsletter
Active	int (0/1)	Cờ thể hiện khách đang “active” trong hệ thống hay không

Chi tiết bài toán

Schema articles.csv

Cột	Kiểu dữ liệu	Ý nghĩa
article_id	int	ID sản phẩm (khóa chính, join với transactions_train)
product_code	int	Mã sản phẩm gốc (nhiều article_id có thể thuộc cùng product_code)
product_type_no / product_type_name	int / str	Loại sản phẩm (áo, quần, váy, phụ kiện, …)
product_group_name	str	Nhóm sản phẩm lớn (Ví dụ: Garment Upper body, Accessories, …)
graphical_appearance_no / graphical_appearance_name	int / str	Kiểu/hoạ tiết (in hoa, trơn, kẻ sọc, …)
colour_group_code / colour_group_name	int / str	Nhóm màu (đen, trắng, pastel, …)
perceived_colour_value_id / perceived_colour_value_name	int / str	Độ đậm nhạt màu (light, dark, intense, …)
perceived_colour_master_id / perceived_colour_master_name	int / str	Nhóm màu chính (blue, red, green, …)
department_no / department_name	int / str	Phòng/bộ phận thời trang (ví dụ: Ladieswear, Menswear, Kids, …)
index_code / index_name	str / str	Index bộ sưu tập / dòng sản phẩm (ví dụ: Divided, H&M Man, …)
index_group_no / index_group_name	int / str	Nhóm index lớn (General, Young, Kids, &c.)
section_no / section_name	int / str	Phân khu chi tiết hơn trong department
garment_group_no / garment_group_name	int / str	Nhóm loại trang phục (T-Shirts, Dresses, Trousers, …)
detail_desc	str	Mô tả chi tiết sản phẩm (text, dùng cho NLP / embedding nội dung)

Candidate retrieval

Personal rule:

- Order History

  Ý tưởng: lấy item mà user vừa mua gần đây

- Item Pair:

  Ý tưởng: "Customers who bought this also bought that" với cặp sản phẩm item-item

- ALS (Collaborative Filtering- matrix factorization)

  Ý tưởng: dùng **implicit ALS** trên ma trận user–item.

Candidate retrieval

- User Group Time History

  Ý tưởng: top popular items **trong mỗi nhóm user**.

- Most Popular Purchased Items:

  Ý tưởng: top item bán chạy nhất **trên toàn bộ users**.

Feature engineering cho ranker

User features

Feature	Mô tả	Cách tính
age	Tuổi của user	Từ metadata, fillna(0)
user_gender	Giới tính user (0/1/2)	user_gender = 1 nếu (số lần mua đồ nam / tổng số lần mua) ≥ 0.8 user_gender = 2 nếu (số lần mua đồ nữ / tổng số lần mua) ≥ 0.8
FN	Fashion News	Label encoded
Active	Trạng thái hoạt động	Label encoded
club_member_status	Trạng thái thành viên	Label encoded
fashion_news_frequency	Tần suất nhận tin	Label encoded

Feature engineering cho ranker

Item features

Feature	Mô tả	Cách tính
article_gender	Giới tính sản phẩm (0/1/2)	0:Không xác đinh 1:Nam 2:Nữ
season_type	Mùa (0/1/2)	0: không xác định 1: summer (áo phông, quần short...) 2: winter (áo khoác, khăn...)
product_type_no	Loại sản phẩm	Label encoded
product_group_name	Nhóm sản phẩm	Label encoded
graphical_appearance_no	Họa tiết	Label encoded
colour_group_code	Mã nhóm màu	Label encoded
perceived_colour_value_id	Giá trị màu	Label encoded
perceived_colour_master_id	Màu chủ đạo	Label encoded
department_no	Phòng ban	Label encoded
index_code	Mã index	Label encoded
index_group_no	Nhóm index	Label encoded
section_no	Section	Label encoded
garment_group_no	Nhóm garment	Label encoded

Feature engineering cho ranker

Interactive features

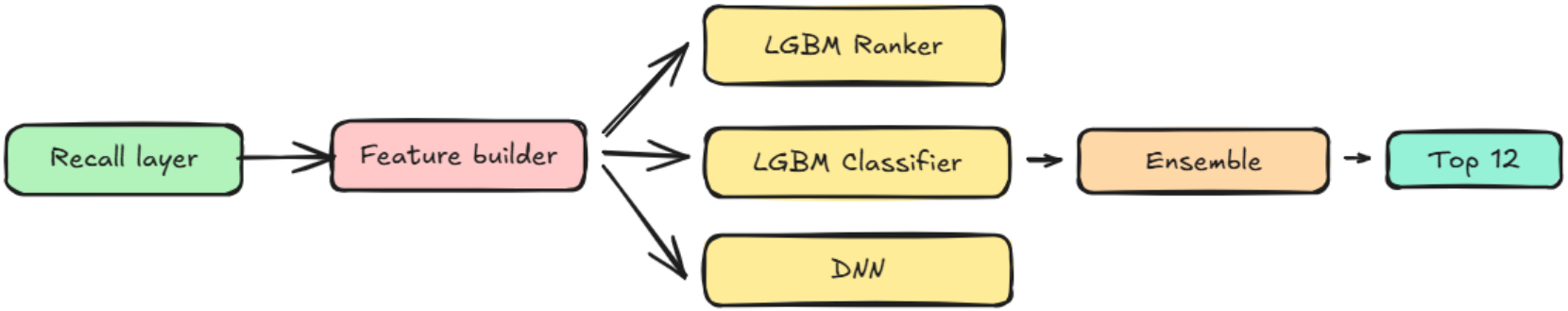
FEATURE	CÔNG THỨC	Ý NGHĨA
full_sale	Tổng số lần bán tích lũy	Độ phổ biến tổng thể
week_sale	Số lần bán trong tuần	Xu hướng gần đây
period_sale	Số lần bán trong N ngày	Hoạt động trong khoảng thời gian
repurchase_ratio	Số user mua lại / Tổng user	Tỷ lệ mua lại
popularity	Σ(1 / (days + 1))	Độ phổ biến có trọng số thời gian

Feature engineering cho ranker

Embedding Similarities features

Feature	Embedding Source	Dimension	Cách tính	Ý nghĩa
dssm_similarity	DSSM model	128	dot_product(user_emb, item_emb)	Độ tương đồng từ metadata và interaction
yt_similarity	YouTube DNN	128	dot_product(user_emb, item_emb)	Độ tương đồng từ sequence patterns
wv_similarity	Word2Vec	128	dot_product(user_emb, item_emb)	Độ tương đồng từ co-occurrence

Kiến trúc tầng Ranking & Ensemble



Input: candidates(customer_id, article_id)+feature (user /item /dynamic /CF / embedding)

3 Model: LGBM Ranker, LGBM Classifier, DNN

Output cuối: Top-12 article cho mỗi user (ensemble)

Model 1: LGBM Ranker (LambdaRank)

- **Nhóm theo user (query):**

Mỗi customer_id = 1 query với danh sách candidates.

- **Loss:** LambdaRank / NDCG@12

gradient dựa trên cặp (i, j) cùng user, label khác nhau, tối ưu ΔNDCG

- **Train:**

dùng toàn bộ feature tabular

tree-based → không cần chuẩn hóa phức tạp, chú ý dtype để tiết kiệm RAM

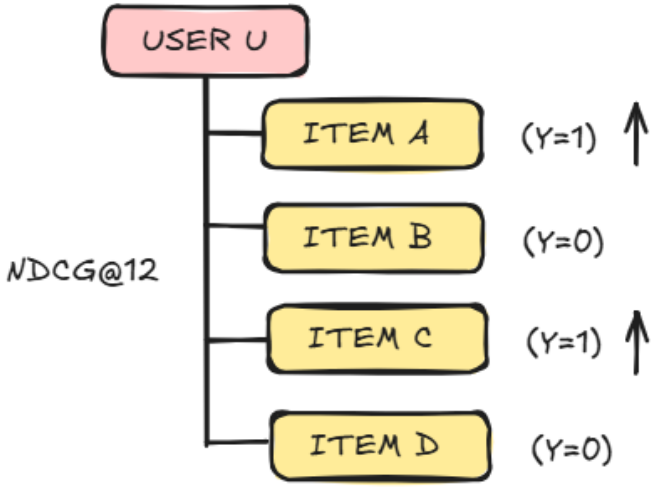
- **Inference:**

predict score cho từng (u,i)

sort theo score, lấy Top-12 / user

- **Lý do chọn:**

tối ưu trực tiếp thứ hạng, mạnh với tabular, train nhanh, single-model score cao



"LambdaRank học sao cho item y=1 đứng trên item y=0 trong cùng user".

Model 2: LGBM Classifier (Binary CE)

- **Đơn vị mẫu:** mỗi (u,i) là 1 sample, **không group**
- **Loss:** BCE :

$$L = -[\, y \log p + (1-y) \log(1-p) \,]$$

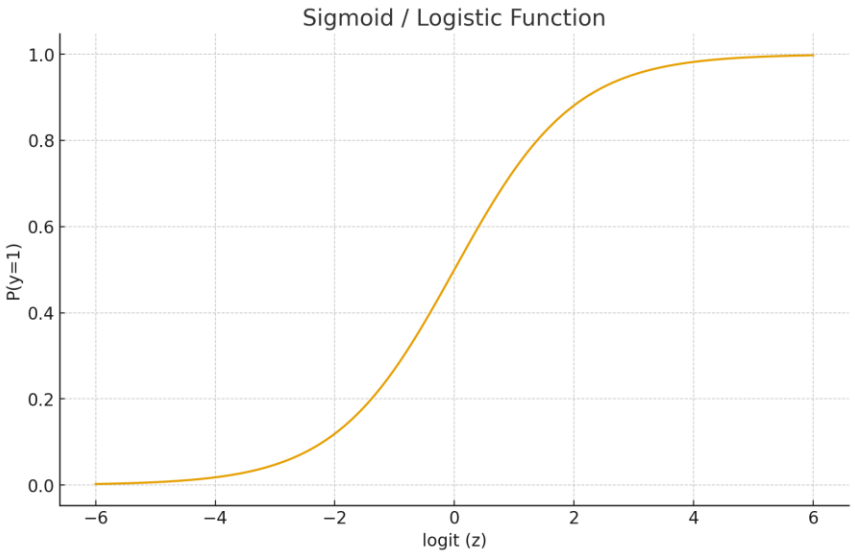
thường dùng scale_pos_weight để xử lý imbalance

- **Train:** dùng cùng bộ feature, số cây ít hơn Ranker (vd 300–700)
- **Inference:**

predict xác suất p = P(buy | u,i)
dùng p để xếp hạng (sort top-12)

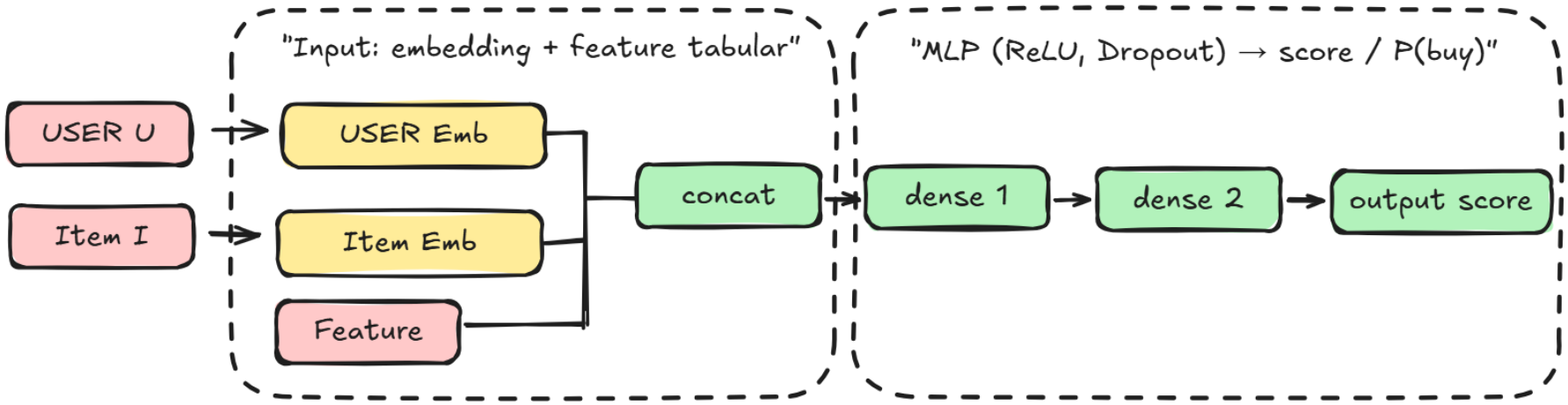
- **Lý do chọn:**

training ổn định, dễ tuning, góc nhìn "probability" khác Ranker → tăng diversity khi ensemble



	LGBM Ranker	LGBM Classifier
Train unit	user-level list	(u,i) độc lập
Loss	LambdaRank (NDCG)	BCE
Output	ranking score	probability p

Model 3: DNN (Embedding + MLP)



- **Kiến trúc:**

Embedding **user/item** (pretrain + fine-tune); có thể embed 1 số cat feature
Concat: [user emb, item emb, dynamic features, CF scores, recency,…]
MLP 2–3 tầng: ReLU, Dropout, BatchNorm (tùy notebook)

- **Loss:**

BCE or pairwise logistic / hinge
in-batch negative sampling để học tốt hơn

- **Hyper đang sử dụng**

hidden 128–256, dropout 0.1–0.3, batch 4k–16k, Adam lr 1e-3 → 1e-4, 3–8 epochs (early stopping)

- **Inference:**

tính embedding + forward MLP → score, sort top-12 per user

Ensemble

- Mỗi bài toán gợi ý, mỗi mô hình (LightGBM, LightGBM, DNN) sẽ đưa ra điểm số dự đoán cho từng ứng viên.
- Để tận dụng ưu điểm của từng mô hình, ta kết hợp chúng bằng cách **ensemble** – tức là lấy trung bình có trọng số các dự đoán của từng mô hình .
- Công thức:

$$\text{FinalScore} = \alpha \cdot \text{LGB\_RScore} + \beta \cdot \text{LGB\_CScore} + \gamma \cdot \text{DNNScore}$$

Trong đó:

- α, β, γ là các trọng số được tinh chỉnh trên tập validation để tối ưu chỉ số MAP@12.
- Các trọng số này phản ánh mức độ tin cậy của từng mô hình.

HUST

hust.edu.vn fb.com/dhbkhn

THANK YOU !