viettel
Theo cách của bạn

VHAC 2023 - News Recommendation
1st place solution

Team DHA

Ngô Hoàng Đăng - Trung tâm Không gian mạng Viettel
Nguyễn Huy Hoàng - Trung tâm Không gian mạng Viettel
Nguyễn Huy Anh - Trung tâm Không gian mạng Viettel

viettel
cyberspace

PHẦN 1

XU HƯỚNG HỆ THỐNG ĐỀ XUẤT TRÊN THẾ GIỚI

viettel
cyberspace

1.1 Tổng quan về hệ thống đề xuất

Hệ thống đề xuất là một ứng dụng lọc thông tin được hình thành từ việc học các sở thích và tương tác của người dùng nhằm có thể đề xuất nội dung phù hợp với họ.

Mục tiêu chính của hệ thống đề xuất là **cung cấp những gợi ý cá nhân hóa và tối ưu hóa trải nghiệm người dùng,** giúp họ **khám phá được nội dung, sản phẩm hoặc thông tin mới** mà họ có thể chưa biết hoặc chưa nhìn thấy.

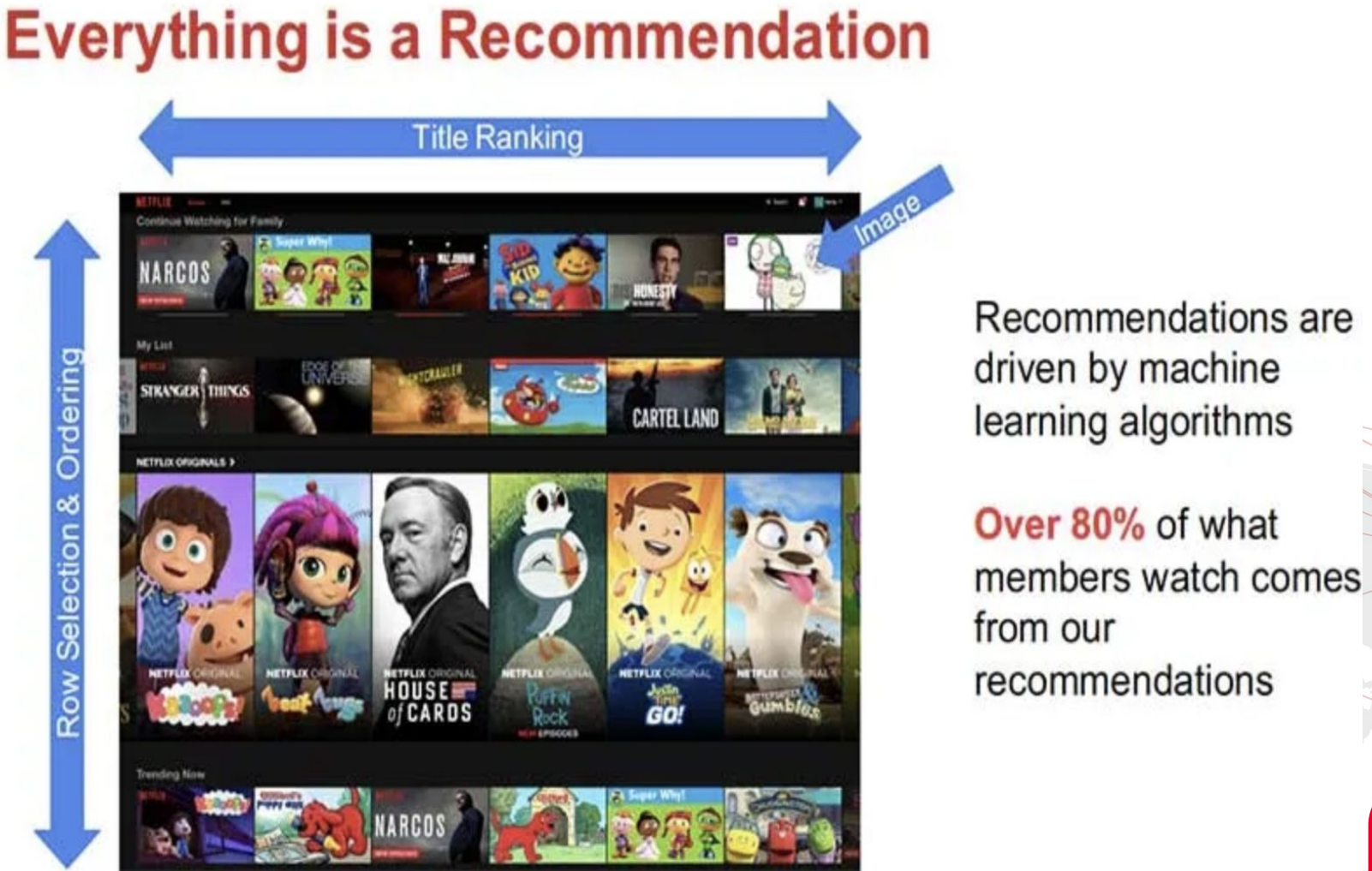
viettel
cyberspace

1.2 Lợi ích của hệ thống đề xuất

Hệ thống gợi ý thúc đẩy sự tương tác của khách hàng và doanh thu.

NETFLIX

- Hệ thống đề xuất mang đến **80% lượt xem**
- Hệ thống đề xuất **tiết kiệm 1 tỷ đô la mỗi năm**



viettel
cyberspace

1.3 Mục tiêu của track VHAC 2023 - News Recommendation

Thực tế: Hàng chục nghìn bài báo mới được đăng tải hàng ngày, việc đọc hết các nội dung tin bài là **không thể**.

Bài toán: Xây dựng một hệ thống gợi ý tin tức nhằm giảm bớt vấn đề quá tải thông tin và đề xuất các mục tin tức cá nhân hoá cho từng người đọc.

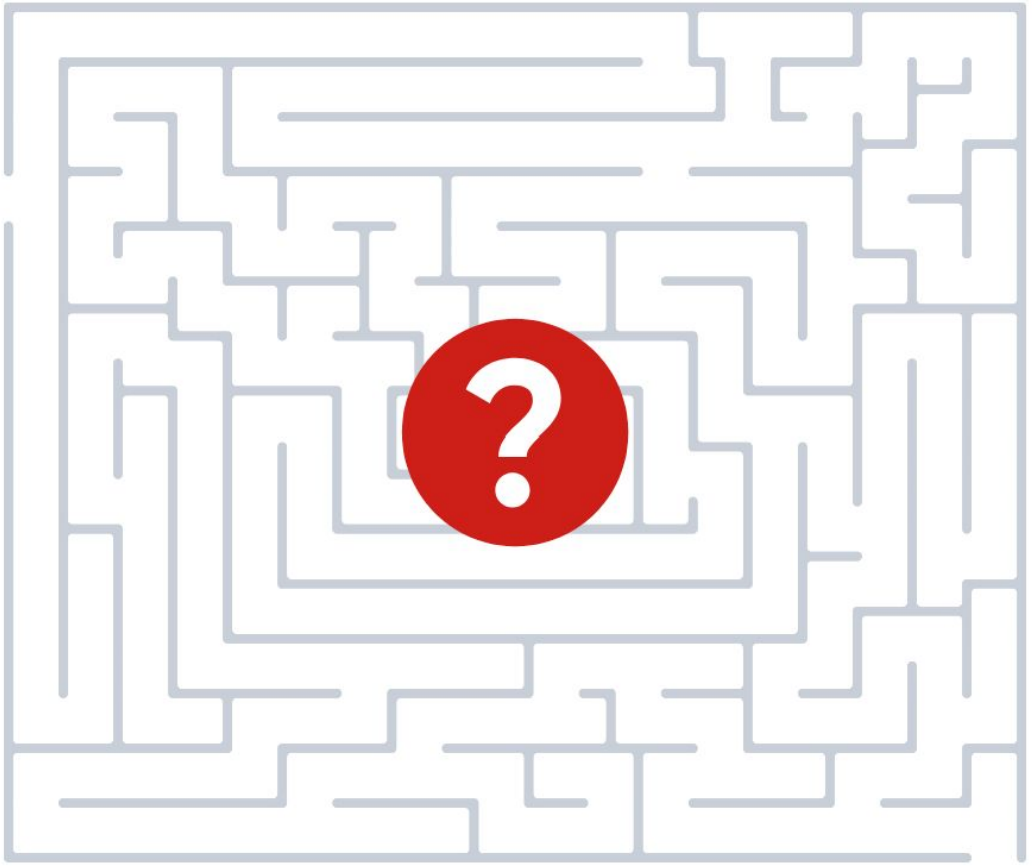
viettel
cyberspace

1.4 Thách thức của track News Recommendation

Vấn đề phụ thuộc dữ liệu: Hệ thống đề xuất yêu cầu dữ liệu đầy đủ và chất lượng để có thể đưa ra gợi ý chính xác. Nếu dữ liệu đầu vào không đủ hoặc không đại diện cho sự đa dạng của người dùng, hệ thống có thể gặp khó khăn trong việc đưa ra gợi ý phù hợp, bao gồm:

Cold-start problem: Khi hệ thống đề xuất gặp người dùng mới sản phẩm mới hoặc hệ thống mới, không có đủ thông tin để xây dựng một mô hình gợi ý đáng tin cậy. Điều này có thể dẫn đến việc hiển thị gợi ý không chính xác hoặc không đủ phù hợp cho người dùng mới.

Tỷ lệ nhiễu cao: Do đặc tính của recommendation báo chí, có những hành vi nhiễu lớn (Người đọc chưa từng xem thể loại trước đây nhưng lại đọc trong các tập valid, private, public test… hoặc có thể do click nhầm).



viettel
cyberspace

1.4 Thách thức của track News Recommendation

	Số tương tác	Số người dùng	Số bài báo	min event time	max event time
Train	10,120,940	88,933	67,107	2022-04-05 23:50:04	2022-04-11 23:59:59
Private test	1,728,018	63,085	27,032	2022-04-12 00:00:00	2022-04-13 00:09:30

cold users: 816
cold items: 8,063

interactions with cold users: 48,448
interactions with cold items: 572,890

interactions with cold users or cold items: 606,262 (35%)
interactions with cold users and cold items: 15,076

viettel
cyberspace

PHẦN
2

GIẢI PHÁP CỦA NHÓM

viettel
cyberspace

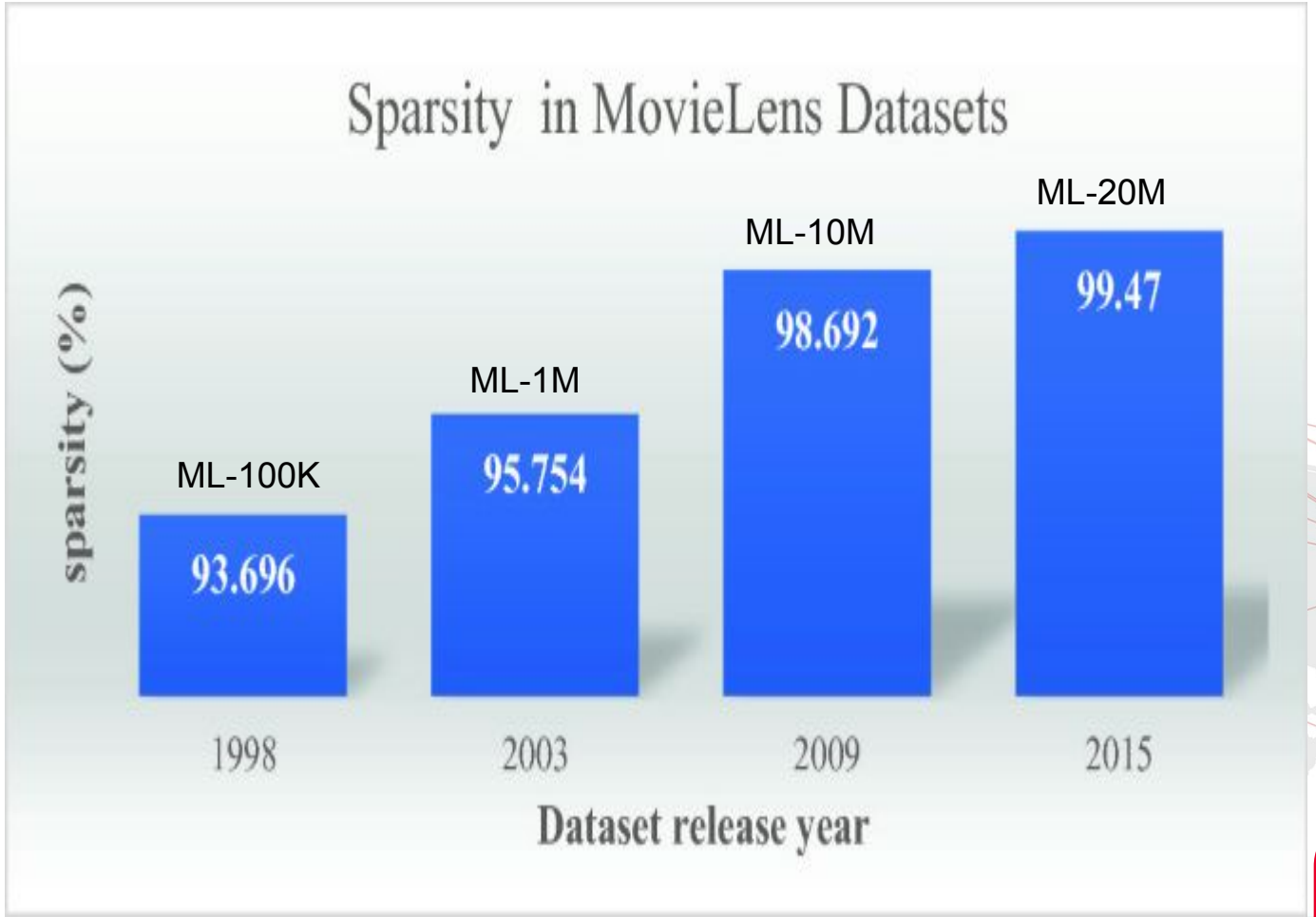
2.1 Các hướng tiếp cận với bài toán đề xuất

2.1.1 Hướng tiếp cận sử dụng phương pháp phân rã ma trận (Matrix factorization)

Không là phương pháp tối ưu cho bài toán:

- Cold Start Problem: 35% lượng tương tác đọc bài báo ở ngày t + 1 là giữa người dùng với bài báo mới được đăng tải.

- Sparsity: 99,83%



viettel
cyberspace

2.1 Các hướng tiếp cận với bài toán đề xuất

2.1.2 Hướng tiếp cận sử dụng mạng nơ-ron đồ thị (Graph neural network)

Nhóm thử nghiệm đánh giá thử nghiệm quét với 15 phút một phiên. Tuy nhiên độ dài của một chuỗi tương tác thấp (độ dài trung bình <3 bài báo một chuỗi).

Nguyên nhân dự đoán:

- Việc đề xuất là do dạng impression ngẫu nhiên (push-app) không hẳn là xem bài viết tiếp theo nội dung của nội dung đang đọc.

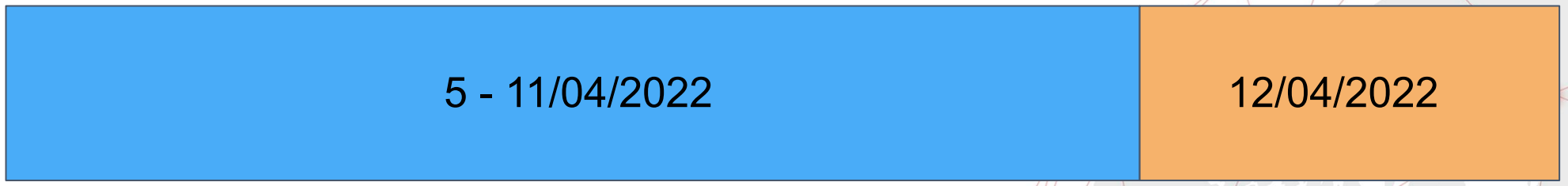
viettel
cyberspace

2.2 Hướng tiếp cận của nhóm

How to track leaderboard precisely?

Local validation

Make submission



viettel
cyberspace

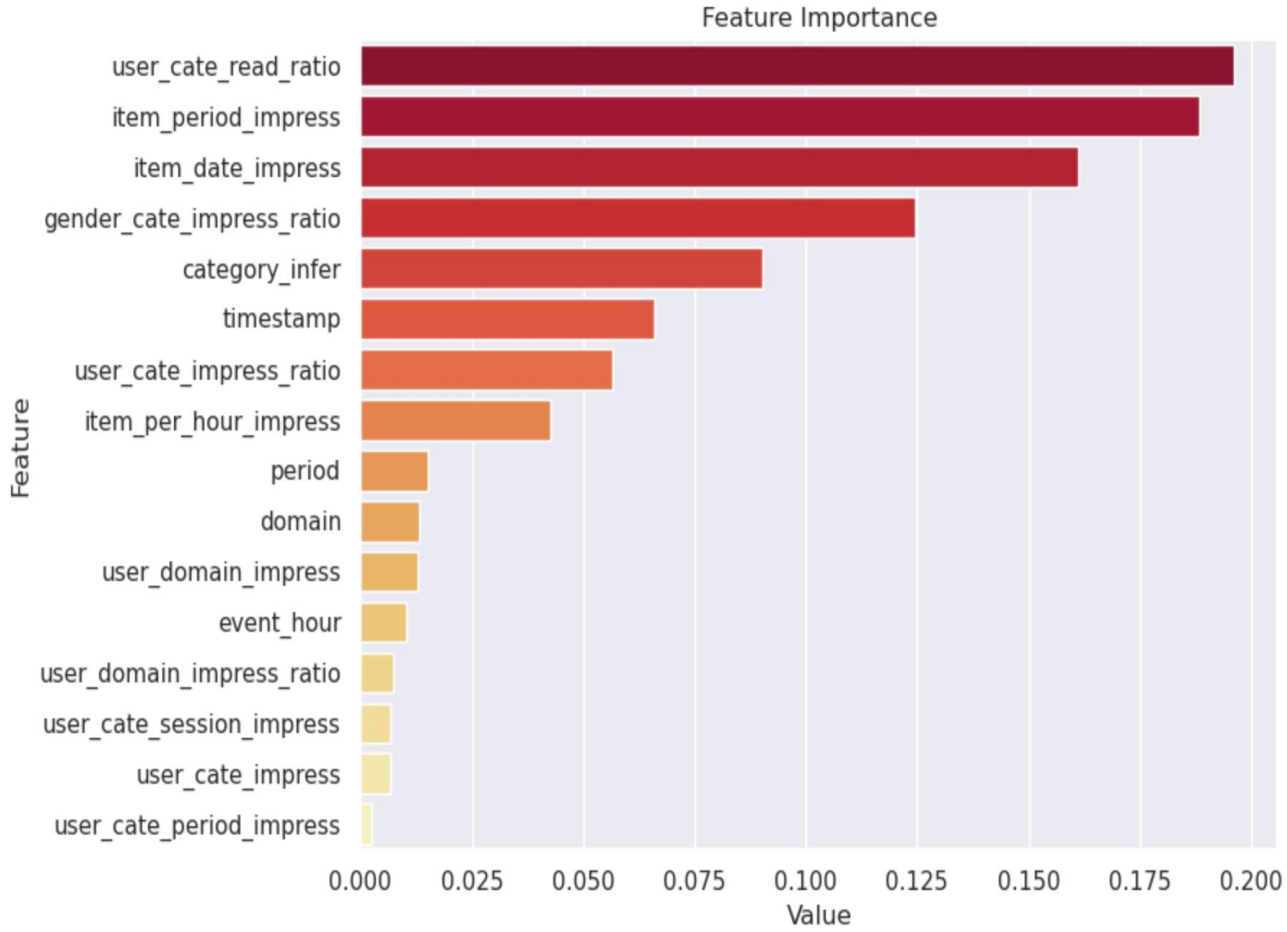
2.2 Hướng tiếp cận của nhóm

Xây dựng đặc trưng (Feature engineering) ~ create hundreds of features

user features: user impression, user engagement ratio.

item features: title text embedding, domain, category, item-date impression, item-period impression, item-hour impression.

user-item interaction features: user-category read ratio, time since article was published, event period (morning, afternoon, evening, night), gender-category impression ratio.



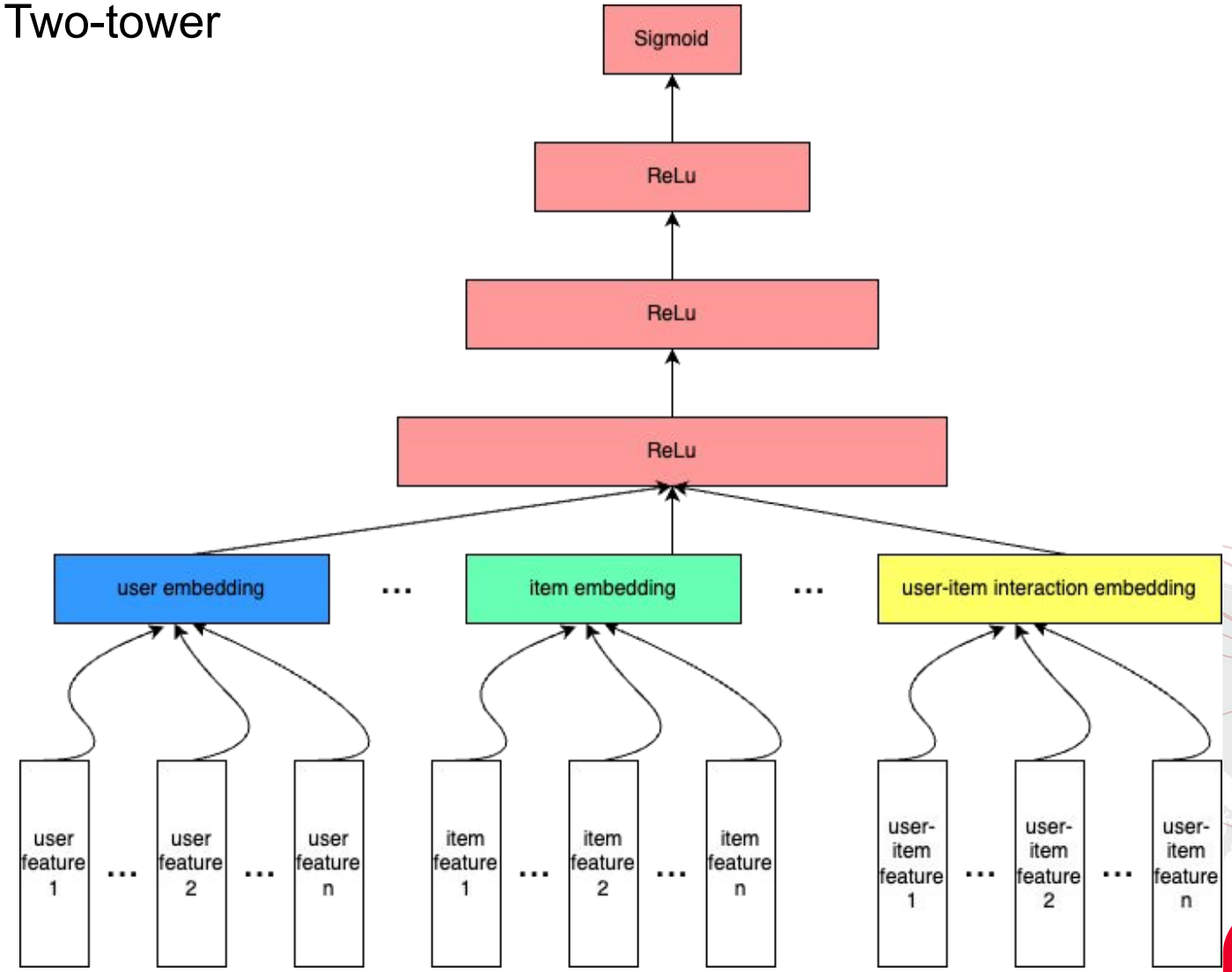
viettel
cyberspace

2.2 Hướng tiếp cận của nhóm

Hướng tiếp cận sử dụng mô hình gợi ý Two-tower

Tích hợp thông tin người dùng/ sản phẩm (Side-information): Thông tin chi tiết về người dùng/ sản phẩm được tích hợp vào trong mô hình, do đó hệ thống có thể giải quyết vấn đề về cold-start của các bài báo mới.

Khả năng học đặc trưng phức tạp: Two-Tower sử dụng mạng nơron cho cả người dùng và sản phẩm để học các đặc trưng, thông tin ẩn, ngữ cảnh ẩn trên tương tác người dùng và bài báo. So với phương pháp sử dụng matrix factorization, mô hình Two-Tower có khả năng học các biểu diễn phức tạp hơn.



Tham khảo:
P Covington, Deep Neural Networks for YouTube Recommendations

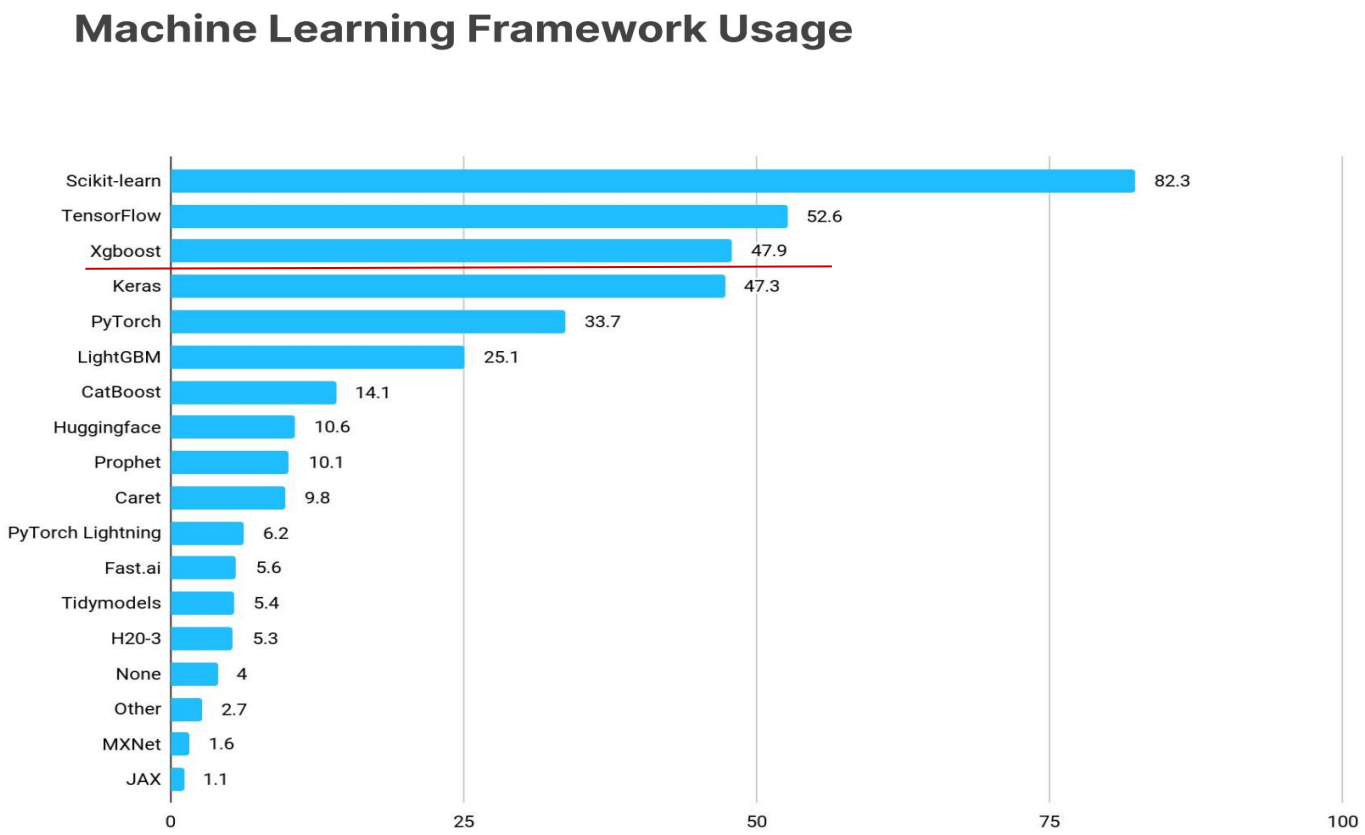
viettel
cyberspace

2.2 Hướng tiếp cận của nhóm

Hướng tiếp cận sử dụng lớp mô hình gradient-boosted trees - XGBoost

Dễ dàng sử dụng và triển khai trong thực tế: Có sẵn nhiều thư viện và framework hỗ trợ GBDT như XGBoost, LightGBM và CatBoost, giúp đơn giản hóa việc triển khai và sử dụng.

Tích hợp thông tin sản phẩm: GBDT có thể tích hợp thông tin chi tiết về sản phẩm, bao gồm các đặc trưng sản phẩm (ví dụ: miêu tả sản phẩm, danh mục, thương hiệu) để cải thiện khả năng gợi ý.



Kaggle | State of ML & Data Science 2021

viettel
cyberspace

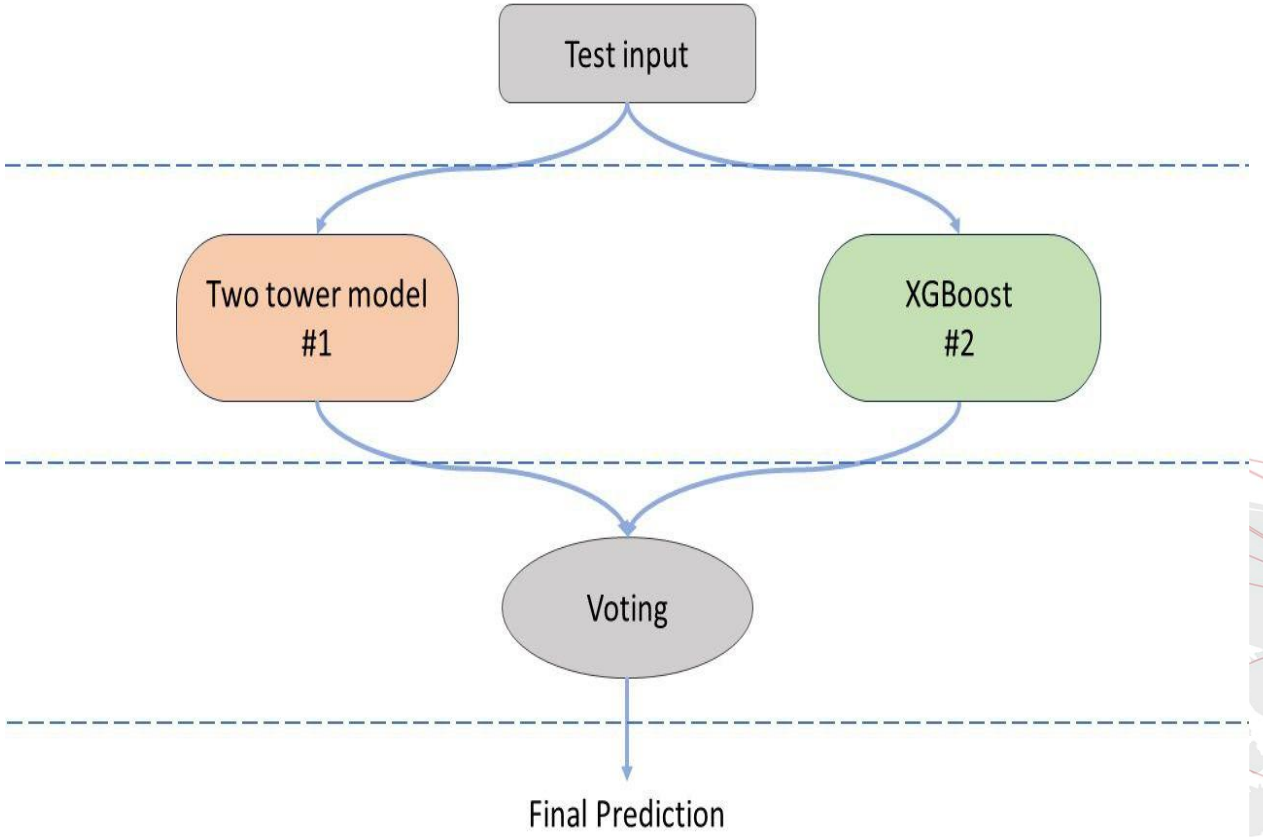
2.2 Hướng tiếp cận của nhóm

Hướng tiếp cận sử dụng kỹ thuật ensemble learning

- Kết hợp ưu điểm của cả hai phương pháp sử dụng deep learning và gradient-boosted tree
- Hiệu suất dự đoán tốt hơn

Thí nghiệm của nhóm:

- Simple averaging
- Power averaging
- Hill climbing***



viettel
cyberspace

PHẦN
3

KẾT QUẢ ĐẠT ĐƯỢC

Kết quả thực nghiệm

Model	Local validation	Private test
Matrix factorization	0.80732	0.80747
Two tower (with side information)	0.98876	0.98890
XGBoost (with side information)	0.98828	0.98855
Two tower + XGBoost	0.98946	0.98957

viettel
cyberspace

PHẦN
4

KẾT LUẬN

viettel
cyberspace

Giá trị đem lại của hệ thống đề xuất

Nhóm nghiên cứu đã thử nghiệm một số giải pháp đánh giá khả năng đọc của người đọc với các bài báo qua nhiều phương án. Các phương án đều cho các kết quả khả quan trong việc bao quát hóa bài toán. Để giải quyết các vấn đề như cold-start, click-noise, clickbait…, việc kết hợp giữa các mô hình đem lại các kết quả tốt nhất trong các giải pháp trên.

Dựa trên các thành tựu đã đạt được trong quá trình nghiên cứu và kết quả tham gia cuộc thi xây dựng hệ thống đề xuất cho track News Recommendation, nhóm nghiên cứu nhận thấy lợi ích của việc đưa hệ thống tương tự cho các sản phẩm báo đề xuất của tập đoàn như Tiin.vn nhằm nâng cao chất lượng, trải nghiệm khách hàng hay Viettel Family nhằm tăng tương tác của cán bộ nhân viên với các sự kiện, thông tin của Tập đoàn.

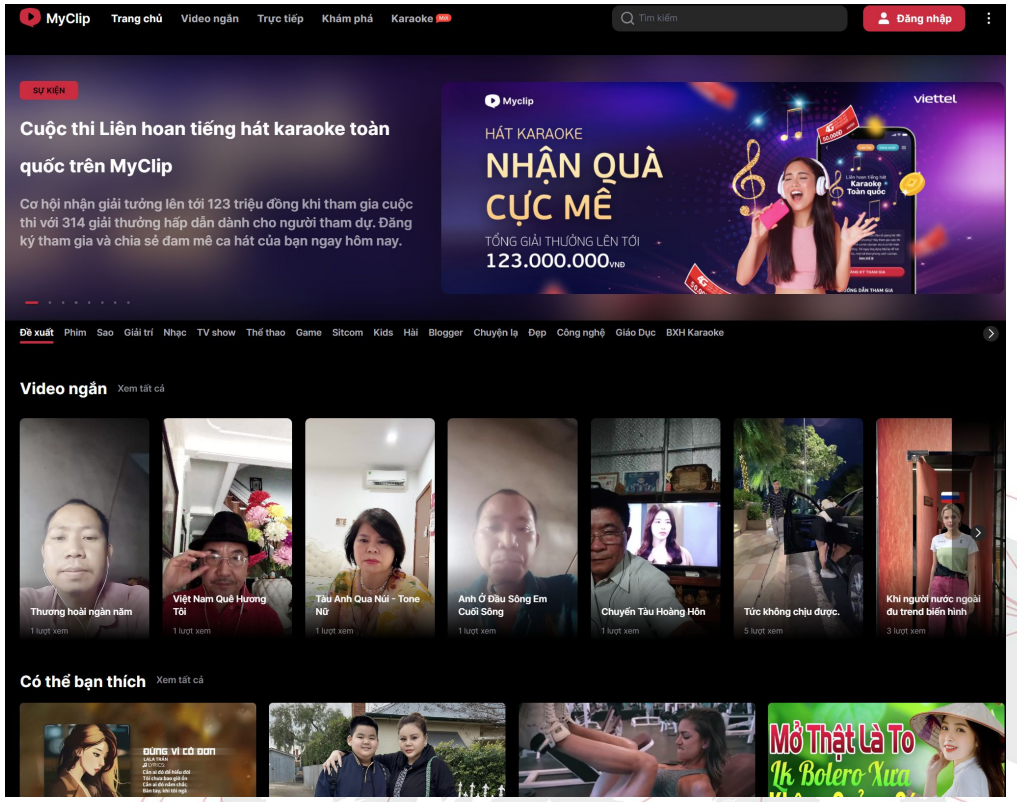
viettel
cyberspace

Mô hình hệ thống đề xuất thiết kế cho nền tảng Myclip

Hệ thống đề xuất được nghiên cứu và phát triển bởi đội ngũ kỹ sư VTCC nhằm giải quyết các yêu cầu phức tạp cho hệ thống Myclip:

Thách thức:

1) Hệ thống Myclip là nền tảng chia sẻ số lượng nội dung lớn với **3,5 triệu nội dung và 1,5 triệu khách hàng.**
2) Hệ thống Myclip có mức độ tương tác không đồng đều giữa các nội dung, có nội dung rất nhiều tương tác, nội dung ít tương tác…
3) Là hệ thống chia sẻ Video, nên thời gian phản hồi của hệ thống đề xuất rất thấp < 0.1ms.
4) Dữ liệu được lưu trữ với mục đích vận hành, nên chưa được chuẩn hóa theo các nhu cầu thiết kế hệ thống đề xuất…



Kết quả:

1) Việc áp dụng các mô hình kỹ thuật đề xuất và thiết kế hệ thống giúp đạt được **15-35% Click-through rate** **(tăng 8% so với hệ thống cũ)** với **0.058ms tốc độ phản hồi** cho toàn bộ tải hệ thống. Sẵn sàng mở rộng cho toàn bộ nhu cầu phát triển của nền tảng Myclip.
2) Hệ thống tự động chuẩn hóa, điều chỉnh dữ liệu cho các video.
3) Xây dựng cơ sở hành vi khách hàng, hành vi tương tác video, cơ sở lưu logs… nhằm giúp cho việc phân tích chuyên sâu về khách hàng, tạo tiền đề cho việc xây dựng các sản phẩm mới có tính cạnh tranh trên thị trường.

viettel
cyberspace

Slide tham khảo

2.2.3 Hướng tiếp cận sử dụng phương pháp Ensemble learning

Power averaging:

Choose a power p = 2, 4, 8, 16

PowerAverage(p) = (Pred1^p + Pred2^p + … + Predn^p) / n

Simple averaging:

SimpleAverage = 0.5 * Pred1 + 0.5 * Pred2

Hill climbing: iteratively finding the best next model and blending via a linear combination with the current best predictions.

viettel
cyberspace

Slide tham khảo

Xây dựng vector biểu diễn cho tiêu đề bài báo (Sentence embedding)

Các phương pháp nhóm đã thử nghiệm:

- Sentence Transformer
- PHO Word2Vec