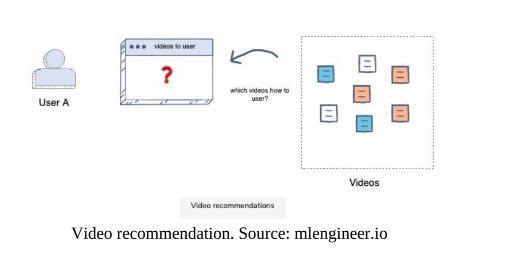
**Seminar :Youtube Video Recommondations**

**I) Problem Statement.**

- Xây dựng đề xuất video cho người dùng YouTube

-Tác giả muốn tối đa hóa mức độ tương tác của người dùng cũng như đề xuất các loại nội dung mới cho người dùng.



**II)** **Metrics design and Requirements**

**a)Metrics**

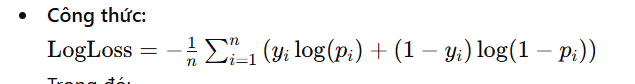
-For offline metrics:Sử dụng

+Độ chính xác – precision. Ví dụ: Trong 10 video truy xuất được, có 8 video đúng liên quan thì Precision = 8/10 = 0.8.

+ recall Ví dụ: Nếu có 20 video liên quan và hệ thống truy xuất được 15 video, thì Recall = 15/20 = 0.75.

+ ranking loss: Ví dụ: Nếu có 100 cặp mẫu và 10 cặp sai thứ tự, thì Ranking Loss = 10/100 = 0.1.

+ logloss: Là hàm mất mát đo mức độ dự đoán xác suất gần với nhãn thực tế.



-For online metrics:Sử dụng:

+ A/B Testing- So sánh hai thuật toán hoặc hệ thống đề xuất video: so sánh tỉ lệ nhấp chuột(CTR),thời gian xem và tỷ lệ chuyển đổi.

**b)** **Requirements**

**\*Trainning:**

+ Hành vi của người dùng thường không thể đoán trước, và video có thể trở nên lan truyền chỉ sau một đêm.

+ Lý tưởng nhất là tác giả muốn huấn luyện mô hình nhiều lần trong ngày để nắm bắt được những thay đổi theo thời gian.

**\*Inference**

+ Đề xuất 100 video cho mỗi người dùng truy cập trang chủ với độ trễ dưới 200ms, lý tưởng là dưới 100ms

+ Đối với các đề xuất trực tuyến, cần cân bằng giữa khám phá và khai thác. Nếu khai thác quá nhiều dữ liệu cũ, video mới sẽ khó tiếp cận. Mục tiêu là cân bằng giữa mức độ liên quan và nội dung mới.

**\*Summary:**

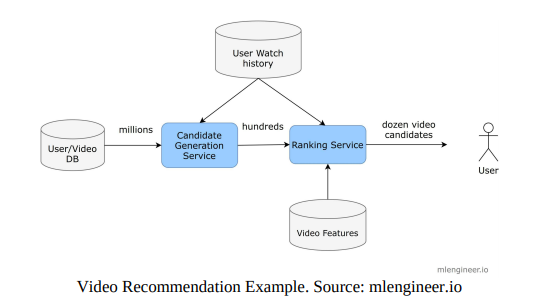
+ Metrics: Độ chính xác hợp lý và recall cao.

+ Huấn luyện với thông lượng cao với khả năng tái huấn luyện nhiều lần trong ngày.

+Độ trễ phải thấp từ 100ms đến 200ms.

+Tìm cách để điều chỉnh cân bằng giữa khám phá và khai thác trong quá trình suy luận.

**III) Multistage Models**

****

- Có 2 giai đoạn: tạo ứng viên(candidate generation) và xếp hạng(ranking service) giúp mở rộng hệ thống:

+ The candidate model: sẽ tìm các video liên quan dựa trên lịch sử xem của người dùng và loại video mà họ đã xem.

+ The ranking model: sẽ tối ưu hóa khả năng xem( video xếp hạng cao sẽ có khả năng xem cao). Phù hợp với thuật toán logistic regression.

**IV) Model training**

**a)Candidate generation model**

-Training data: Tạo training data bằng không gian người dùng - video (user-video watch space).

**+**Chọn khoảng thời gian dữ liệu :tháng trước, 6 tháng gần đây,… giúp tìm ra sự cân bằng giữa thời gian huấn luyện và độ chính xác của mô hình.

- Feature Enginneering: Mỗi người dùng có một danh sách các video đã xem (video, số phút đã xem).

- Model:

+Có thể thực hiện bằng phương pháp phân tích ma trận (matrix factorization).

+Mục đích: tìm nội dung tương đối liên quan dựa trên lịch sử xem của người dùng.

+ Danh sách ứng viên cần đủ lớn để nắm bắt các kết quả tiềm năng và đảm bảo thời gian trễ thấp.

+Collaborative algorithms là một giải pháp hiệu quả do thời gian suy luận nhanh và khả năng nắm bắt sự tương đồng về sở thích giữa người dùng trong user-video space.

**b)** **Ranking model**

-Trong quá trình suy luận, ranking model nhận danh sách video ứng viên từ candidate generation model.

-Với mỗi ứng viên, mô hình ước lượng xác suất video sẽ được xem.

-Sau đó, các video được sắp xếp theo xác suất và trả về cho quy trình phía trên (upstream process).

-Training data:

+Dữ liệu lịch sử xem của người dùng.

+Tỷ lệ xem vs. không xem thường là 2/98=>phần lớn video được đề xuất sẽ không được xem bởi người dùng.

-Features Engineering:

+Watched video Ids: Áp dụng video embedding.

+Historical search query: Áp dụng text embedding (ví dụ: Word2Vec, GloVe, hoặc BERT).

+Training: Tối ưu để xử lý nhanh và có thể huấn luyện lại nhiều lần trong ngày.

+Location: Sử dụng geolocation embedding.

+User associated features: Tuổi, giới tính kèm Normalization hoặc Standardization.

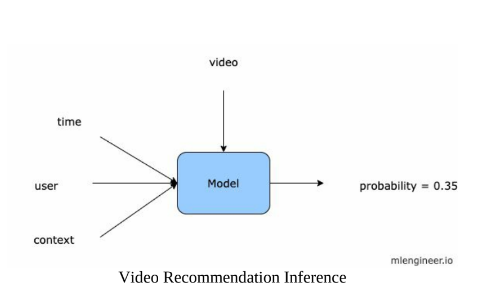
+Aggregated impression: Áp dụng Normalization hoặc Standardization.

+Time-related features: Tháng, tuần trong năm, ngày lễ, thứ trong tuần, giờ trong ngày.

**-**Model:

+Sử dụng DCNv2(Deep & Cross Network v2) với hàm kích hoạt reLu ở lớp ẩn và lớp cuối và Sigmoid.

+Loss funcion là cross-entropy loss.



**V) Calculation and Estimation**

**a)** **Assumptions**

- Số lượt xem video mỗi tháng là 150 tỷ.

-10% số video được xem đến từ các gợi ý.

-Trên trang chính người dùng sẽ có 100 video gọi ý.

-Trung bình sẽ xem 2 video.

-Nếu không nhấp hoặc không xem video trong thời gian nhất định thì video đó bị bỏ lỡ.

-Tổng người dùng là 1.3 tỷ.

-Dữ liệu thu thập trong 1 tháng: 15 tỷ nhãn dương và 750 tỷ nhãn âm.

-Mỗi hàng dữ liệu chứa hàng trăm đặc trưng (~500 bytes).

-Tổng số hàng: 800 tỷ → Dung lượng dữ liệu: 4 Petabytes/tháng.

-Chiến lược lưu trữ: Giữ dữ liệu 6 tháng - 1 năm trong data lake, lưu trữ lâu dài trong cold storage để tiết kiệm chi phí.

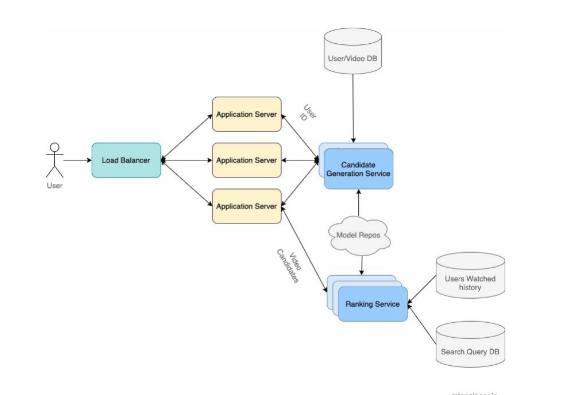
**b)** **Bandwidth and Scale**

-Số yêu cầu gợi ý mỗi giây: 10 triệu người dùng

-Mỗi yêu cầu xếp hạng từ 1.000 - 10.000 video hỗ trợ 1.3 tỷ người dùng.

**VI)System Deign**

**a)Training**

****

-Database:

+User Watched History: Lưu trữ video mà người dùng đã xem theo thời gian.

+Search Query DB: Lưu trữ các truy vấn tìm kiếm mà người dùng đã thực hiện trong quá khứ.

+User/Video DB: Lưu danh sách người dùng và hồ sơ của họ, cùng với metadata video.

+User Historical Recommendations: Lưu trữ các gợi ý đã được gửi đến người dùng trong quá khứ.

-Resampling data: Giảm số lượng mẫu âm để tăng cường quy trình huấn luyện (downsampling negative samples).

-Feature pipeline:

+Pipeline: Chương trình để tạo ra tất cả các đặc trưng cần thiết cho mô hình huấn luyện.

+Yêu cầu thông lượng rất cao vì cần huấn luyện lại mô hình nhiều lần trong vài ngày.

+Công nghệ sử dụng: Có thể dùng Spark (công cụ tính toán trong bộ nhớ), Elastic MapReduce, hoặc Google Dataproc (Spark và Hadoop trên đám mây).

-Model Repos: Lưu trữ tất cả các mô hình lựa chọn phổ biến là AWS S3.

**b)Challenges**

-Kích thước dữ liệu lớn.

-Dữ liệu không cân bằng.

-Độ khả dụng cao.

**c)** **Inference**

-Quy trình yêu cầu gợi ý video từ người dùng:

+Yêu cầu từ người dùng.

+Nhận danh sách ứng viên.

+Ước lượng xác suất xem.

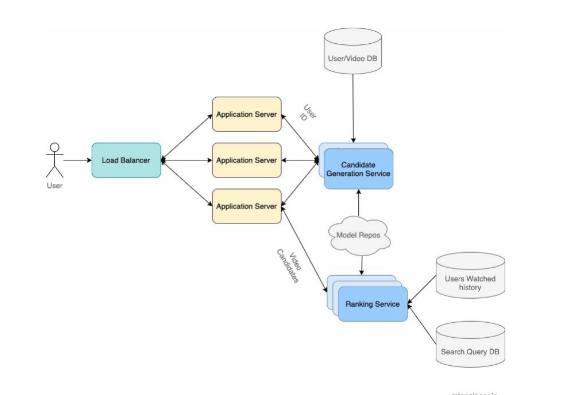
+Trả về video gợi ý.

**VII)** **Scale the Design**

-Mở rộng Application Server.

-Mở rộng dịch vụ candidate generation và ranking

-Có thể sử dụng Kube-proxy để giảm độ trễ.

****