18/12/2024

Quy trình xử lý dữ liệu: bám theo workflow Khôi đã gửi, có thể linh động nếu cần thiết

Nhắc lại câu hỏi:

* Xác định sẽ chỉ dự đoán số lượng stream trên Spotify.
* Loại bỏ các feature liên quan đến các nền tảng khác.

Vấn đề làm về âm nhạc:

* Dự đoán performance của 1 bài hát rất khó, vì nó context-dependent (chủ yếu nằm ở marketing và mức độ nổi tiếng của ca sĩ, vốn không có trong dataset)
* Bài toán ban đầu chỉ dự đoán số lượng streams của Spotify. Đây là một thang đo chủ quan.

**Next step:**

* Khôi: tập trung vào phân tích dữ liệu thô và làm sạch dữ liệu

24/12/2024

Đặt ra vấn đề:

* Các khóa nhạc nên được encode thế nào?

Để ý rằng vòng hợp âm đi theo dạng tròn: đô - đô thăng - rê - rê thăng - mi - fa - fa thăng - sol - sol thăng - la - la thăng - si - đô. Sử dụng vòng tròn lượng giác để encode các nốt nhạc lên vòng tròn lượng giác.

Bổ sung thêm feature ngày ra mắt trong tuần.

Bổ sung thêm câu hỏi:

* Ngày ra mắt ảnh hưởng thế nào đến số lượng stream?
* Khóa chính ảnh hưởng thế nào đến số lượng stream?
* Số lượng ca sĩ tham gia sáng tác ảnh hưởng thế nào đến số lượng stream?
* Tỉ lệ lời bài hát có kéo số lượng stream xuống hay không?

**Next step:**

* Hoàng: làm phần feature engineering và feature selection
* Hoàng: làm phần EDA
* Vân: trả lời các câu hỏi

28/12/2024

Tổng hợp notebook: không có vấn đề

**Next step:** Phân công thuyết trình:

* Khôi: phase 1, phase 2 và phase 3 phần Data Cleaning
* Hoàng: phase 3 phần Data Transformation và phase 4 phần EDA
* Vân: phase 4 phần trả lời câu hỏi và phase 5