1. Warm up questions
2. A machine learning pipeline
3. Problem define: đặt ra bài toán cần giải quyết:

Các ứng dụng công nghệ phần mềm kĩ thuật cao hiện tại đều sinh ra để giải quyết 1 vấn đề nào đó trong hệ thống hoặc giúp cải thiện năng suất lao động, thay thế con người làm những công việc lặp lại.

AI cũng vậy, chúng ta luôn cần phải define 1 bài toán, 1 vấn đề đang gặp phải và liệu có thể sử dụng AI hay ko?

Sẽ có 2 khả năng:

Đã có data: data đã có nhưng có đủ sử dụng cho việc giải quyết bài toán ko?

Chưa có data: Cần thu thập những data thế nào? Dưới format gì, sử dụng database ra sao,..

Đó là các câu hỏi chúng ta luôn cần phải đặt ra trước khi nghĩ đến việc giải quyết 1 bài toán bằng AI

Nếu vấn đề simple và có thể giải quyết dễ dàng bằng rule based -> đừng dùng AI!

1. Data Collection and storage:

Dữ liệu được thu thập theo thời gian bằng các cách như:

* Online content analysis
* Online focus groups
* Online interviews
* Online qualitative research
* Online questionnaires
* Social network analysis

Lưu dữ liệu: Dữ liệu sẽ được lưu vào các database như MongoDB, SQL,.. với 1 số lượng record lớn để dễ dàng biến đổi và sử dụng cho bài toán.

Thông thường dữ liệu được lưu vào MongoBD với dạng các json chứa các record

Ví dụ:

{user:’mr john cena’, phone:01234567, address : “abc abc”, image:’path”}

1. Phân tích dữ liệu:

Sau khi có dữ liệu, vậy việc phân tích dữ liệu để làm gì? Có ảnh hưởng ra sao đến model học máy?

Ví dụ: Trong trường hợp thực tế, khi doanh nghiệp của bạn không phát triển, thì cần biết và hiểu được những sai sót đang móc phải, kể cả khi đang phát triển thì chúng ta cũng cần xác định những chỉ số có thể cải thiện để càng phát triển thêm nữa. Vậy phân tích dữ liệu giúp ta giải thích và cung cấp dữ liệu có ý nghĩa cho việc ứng dụng vào model thuật toán học máy tốt hơn.

Data Cleaning (Làm sạch dữ liệu)

Có những dữ liệu được thu thập có thể không hữu ích hoặc không liên quan đến mục tiêu phân tích của bạn, do đó dữ liệu cần được làm sạch. Dữ liệu được thu thập có thể chứa các bản ghi trùng lặp, khoảng trắng hoặc lỗi. Giai đoạn này phải được thực hiện trước phân tích vì dựa trên việc làm sạch dữ liệu, đầu ra phân tích của bạn sẽ gần hơn với kết quả mong đợi.

Data Analysis (Phân tích dữ liệu)

Sau khi dữ liệu được thu thập, làm sạch và xử lý, nó đã sẵn sàng để phân tích. Trong quá trình phân tích bạn có thể thấy thông tin chính xác bạn cần hoặc bạn cần thu thập thêm dữ liệu. Trong giai đoạn này, bạn có thể sử dụng các công cụ và phần mềm phân tích dữ liệu để giúp bạn hiểu, giải thích và rút ra kết luận dựa trên các yêu cầu.

Trực quan hóa dữ liệu thể hiện dữ liệu dưới dạng biểu đồ và đồ thị. Nói cách khác, dữ liệu được hiển thị bằng đồ họa để bộ não con người dễ hiểu và xử lý nó hơn. Trực quan hóa dữ liệu thường được sử dụng để thể hiện các facts (sự thật) và xu hướng chưa biết. Bằng cách quan sát các mối quan hệ và so sánh các bộ dữ liệu, bạn có tìm ra những thông tin có ý nghĩa.

Ví dụ: với số liệu cho thấy người dùng trên các sàn thương mại điện tử tăng lên vào các tháng cuối năm, cũng là các tháng có nhiều event sự kiện trong năm (blackfriday, event sale của các nhãn hàng,…)

Feature engineering:

Sau quá trình phân tích, ta có thể thấy được correlation giữa các thông tin với nhau, những thông tin nào có ích và ý nghĩa với bài toán hiện tại. Những thông tin nào nhiều nhiễu (outlier) và thiếu dữ liệu (missing data) có thể bị loại bỏ,…

Model ML chủ yếu tính toán dựa trên số thực, nên các feature quan trọng ở dạng str, list, .. cần được vector hoá.

Ví dụ chúng ta có 1 cột thông tin là “Đảng phái” gồm các thông tin “Trump voter”, “Biden voter”, “kanye west”,… chúng ta cần đưa thông tin này về dạng số hoá, như trump\_voter = 1, biden\_voter =2 ,…

Sử dụng data để train các model ML

Đánh giá model

Đóng gói model

Deploy

Monitor inference model.