Mình đã dành vài tháng qua để phân tích data từ các cảm biến, các khảo sát và các logs. Dù có tạo bao nhiêu biểu đồ đi chăng nữa, hay thuật toán phức tạp thế nào, kết quả vẫn chỉ là cú lừa.

Ốp một model random forest vào data cũng giống như ta truyền virus vào nó vậy. Virus đó không có mục đích gì khác, ngoài việc làm sai lệch insights, data giống như là một thùng rác đầy vậy.

Thậm chí tệ hơn, khi bồ đưa những phát kiến tới cho CEO xem, và đoán xem nào? "Ủa em?" Anh ấy/ Chị ấy ngửi thấy có gì đó sai sai, phát hiện của bồ không đúng với hiểu biết của họ về lĩnh vực đó – Sau tất cả, họ là chuyên gia về lĩnh vực đó, họ hiểu biết hơn bồ, bồ chỉ là một analyst hoặc một dev thôi.

Ngay sau đó thì, máu dồn lên mặt (không hề lên não), mắt mờ tay run, một thoáng im lặng, và tiếp theo là lời xin lỗi.

Điều đó cũng chưa hẳn là tệ nhất. Điều gì sẽ xảy ra, nếu phát kiến của bồ là sự cam đoan, khẳng định chắc nịch (lúc này chả có ai nhận ra rằng nó sai như bên trên cả), và rồi vận mệnh cả công ty đặt vào nó?

Vậy là bồ đã nuốt một đống data bẩn rồi, bồ bảo với công ty hãy làm gì đó với những kết quả này mà không hề biết nó sai. Bồ sẽ gặp nhiều phiền phức đó.

Data không chính xác hoặc không nhất quán, có thể dẫn tới các kết luận sai lầm. Vì thế, bồ làm sạch và hiểu data bao nhiêu, thì chất lượng của kết quả sẽ tốt bấy nhiêu.

Hai ví dụ trích từ Wikipedia.

> Giả sử, chính phủ muốn phân tích số liệu điều tra dân số, để quyết định, khu vực nào cần chi tiêu và đầu tư nhiều hơn vào cơ sở hạ tầng. Trong trường hợp này, điều quan trọng là cần phải có dữ liệu đáng tin cậy để **tránh các quyết định sai lầm**.
>
> Trong lĩnh vực kinh doanh, dữ liệu không chính xác có thể sẽ phải trả giá đắt. Nhiều công ty sử dụng database thông tin khách hàng, có các thông tin như liên lạc, địa chỉ, sở thích. Giả sử, địa chỉ không chính xác, công ty sẽ bị **thiệt hại chi phí** gửi lại thư hoặc tệ hơn là **mất khách hàng**.

> “Garbage in, garbage out.”

Trên thực tế, một thuật toán đơn giản có thể tốt hơn cái phức tạp chỉ cần nó có dữ liệu đủ và tốt.

> “Quality data beats fancy algorithms.”

Vì những lý do đó, điều quan trọng là cần phải có bảng hướng dẫn từng-bước, một cheat sheet, các bước kiểm tra sẽ được áp dụng.

Nhưng trước hết, cần phải biết được chúng ta đang cố gắng để đạt được cái gì? Nó có ý nghĩa gì với chất lượng dữ liệu? Lấy gì làm thước đo chất lượng dữ liệu?

Index:

Chất lượng dữ liệu

Thẳng thắn mà nói, mình không tìm được một bài giải thích nào tốt hơn cái ở trên Wikipedia. Do đó, mình sẽ tóm tắt nó ở đây.

Có giá trị

Là mức độ data tuân theo các quy tắc kinh doanh hoặc ràng buộc đã xác định trước.

- Ràng buộc Data-Type: giá trị trong các cột cụ thể phải có data-type cụ thể. Ví dụ: boolean, numeric, date,…
- Ràng buộc Range: số và ngày tháng phải trong phạm vi nhất định.
- Ràng buộc Mandatory: một số cột nhất định không được trống.
- Ràng buộc Unique: trường, hoặc tổ hợp các trường (tính năng, cột…) phải unique
- Ràng buộc Set-Membership: giá trị của một cột phải tới từ tập các giá trị rời rạc. Ví dụ giới tính chỉ có thể là Nam hoặc Nữ.
- Regular expression patterns: một số trường phải theo pattern nhất định. Ví dụ số điện thoại (+84) 123456789 hoặc số báo danh của thí sinh thi THPTQG 00000001.
- Ràng buộc Cross-Field: một số điều kiện trải dài trên nhiều trường. Ví dụ: ngày bệnh nhân xuất viện không được sớm hơn ngày nhập viện.

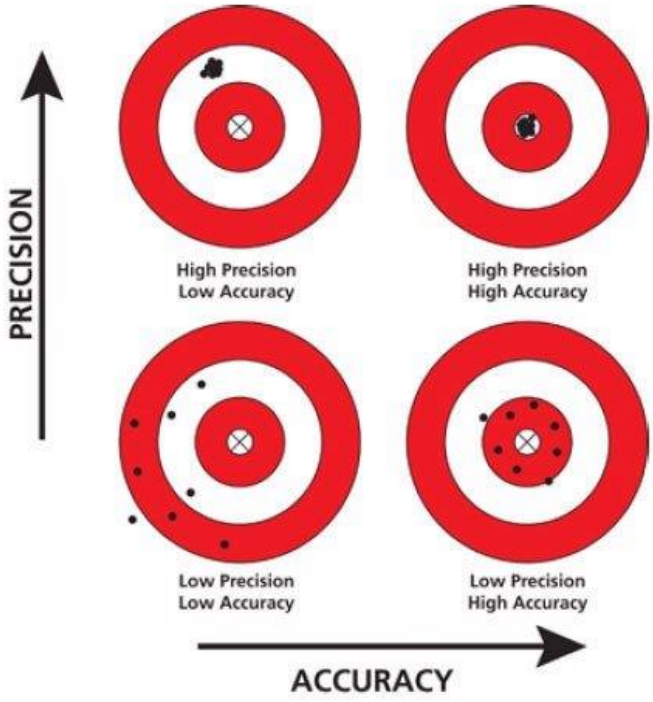
Sự chính xác (Accuracy)

Là mức độ data gần với giá trị thực.

Xác định được các giá trị hợp lệ (tuân theo các ràng buộc), nhưng không có nghĩa rằng chúng đều chính xác (accurate).

Một địa chỉ hợp lệ, nhưng nó lại không tồn tại (Quận 1 là giá trị hợp lệ, nhưng Quận 1, Hải Phòng thì là giá trị không chính xác).

Có một điểm lưu ý giữa accuracy và precision, nếu dịch ra tiếng việt thì đều là độ chính xác. Nhưng tính chính xác thì khác nhau.



Nếu nói rằng bồ sống ở Trái Đất. Nó đúng đấy, nhưng không precise.

Tính hoàn chỉnh

Là mức độ đánh giá tất cả dữ liệu bắt buộc đã biết.

Missing data có thể do nhiều nguyên nhân. Một cách để giảm bớt sự mất mát bằng cách hỏi lại nguồn lấy dữ liệu nếu có thể (đơn vị, người thực hiện, tổ chức khảo sát, thu thập), thu thập lại dữ liệu.

Rất có thể đối tượng khảo sát sẽ đưa ra câu trả lời khác hoặc khó có thể tiếp cận lại họ.

Tính nhất quán

Là mức độ đánh giá sự nhất quán của data, trong cùng một dataset hoặc nhiều datasets.

Sự không nhất quán xảy ra khi hai giá trị trong dataset mâu thuẫn với nhau.

Ví dụ, tuổi là 10, đây là giá trị hợp lệ, nhưng bất hợp lý khi tình trạng hôn nhân lại là Đã ly hôn !?. Hoặc địa chỉ của cùng 1 khách hàng, nhưng trên hai bảng khác nhau, lại có giá trị khác nhau.

Cái nào mới là giá trị chính xác?

Tính đồng nhất

Là mức độ đánh giá data sử dụng cùng một đơn vị đo lường.

Data phải được quy đổi về cùng một đơn vị đo lường, như ngày tháng, cân nặng, chiều cao…

Quy trình

1. Kiểm tra: Xác định các điểm đáng ngờ, không chính xác và không nhất quán.
2. Cleaning: Sửa hoặc loại bỏ các điểm dị thường
3. Verifying: Sau khi cleaning, kết quả phải được xác minh lại.
4.