

XGBoost trong Phân tích Dự báo

Mục tiêu bài học

- Hiểu được **tại sao lại chọn XGBoost** cho bài toán phân tích.
- Nắm rõ **ưu điểm vượt trội của XGBoost** so với các thuật toán khác.
- Biết được **cách thức hoạt động cơ bản của XGBoost** và sự khác biệt giữa hai dạng thuật toán: **tree-based** và **linear**.
- Chuẩn bị nền tảng để bước vào phần trực quan hóa thuật toán XGBoost ở bài học sau.

1. Tại sao chọn XGBoost?

Việc chọn một thuật toán máy học không phải ngẫu nhiên. Nó phụ thuộc vào:

- Bản chất của **vấn đề kinh doanh**.
- Mục tiêu của mô hình là gì (phân loại, hồi quy, v.v).
- Các **yêu cầu phức tạp** và **hạn chế** của dữ liệu.

“ Trong bài toán tiếp thị trực tiếp của chúng ta (telemarketing), chúng ta cần **phân loại nhị phân**: khách hàng sẽ trả lời "Yes" hay "No" với chiến dịch? ”

2. Ưu điểm nổi bật của XGBoost

Ưu điểm	Mô tả
Tầm quan trọng của đặc trưng (feature importance)	XGBoost cho bạn biết đặc trưng nào có ảnh hưởng lớn nhất đến kết quả. Điều này rất hữu ích để điều chỉnh chiến lược kinh doanh.
Độ chính xác cao	XGBoost thường vượt trội hơn so với Logistic Regression, Random Forest, và thậm chí cả Deep Learning trong nhiều trường hợp.
Xử lý song song	XGBoost hỗ trợ xử lý song song ngay trong thư viện — giúp tăng tốc huấn luyện.
Tối ưu lặp lại (iterative learning)	Mô hình học dần từ sai lầm trong từng vòng lặp, cải thiện chính xác dần theo thời gian.

3. XGBoost là gì?

- **Tên đầy đủ:** eXtreme Gradient Boosting.
- **Là thuật toán dạng "ensemble":** Tức là mô hình cuối cùng là sự kết hợp của nhiều mô hình nhỏ (base learners).
- **Sử dụng được cho:**
 - Hồi quy (regression): khi biến mục tiêu là **liên tục**.
 - Phân loại (classification): khi biến mục tiêu là **rời rạc** (như "Yes" / "No").

4. Tree-based vs Linear - Phân biệt 2 cách tiếp cận

Linear (Tuyến tính):

- Phù hợp khi dữ liệu có quan hệ tuyến tính.
- Ví dụ: như đường thẳng trong mô hình hồi quy.

Tree-based (Dạng cây quyết định):

- Dựa vào việc **chia nhánh** theo điều kiện.
- Ví dụ minh họa:
 - Có vỏ bánh không? ☐ → Không phải là bánh pie.
 - Có viền bánh không? ☐ → Là bánh pie.

→ Đây chính là cách mà **thuật toán cây** đưa ra quyết định phân loại. Trực quan, dễ hiểu, và rất hiệu quả với các dữ liệu phức tạp, không tuyến tính.

5. Tại sao chọn tree-based cho XGBoost?

- Cho kết quả chính xác hơn trong thực tế.
- Phù hợp với các dữ liệu đa chiều, nhiều điều kiện phức tạp.
- Thực thi tương tự như linear trong mã nguồn, chỉ cần cấu hình một vài tham số khác biệt.

Kết luận

- XGBoost là công cụ cực mạnh cho cả phân loại và hồi quy.
- Trong trường hợp của chúng ta (phân loại khách hàng trả lời "Yes" hoặc "No"), XGBoost sẽ giúp đưa ra mô hình chính xác và khả thi để áp dụng trong kinh doanh.
- Ở video tiếp theo, chúng ta sẽ **hiểu sâu hơn về trực giác bên trong của thuật toán XGBoost** — với đồ thị và ví dụ minh họa quá trình học của mô hình.

Tác giả: **Đỗ Ngọc Tú**
Công Ty Phần Mềm **VHTSoft**

