## **Workflow cơ bản của một data scientist là gì?**

Để hiểu hơn về workflow (quy trình làm việc) của một Data Scientist là gì, hãy bắt đầu với quy trình làm việc của Blitzstein & Pfister được giảng dạy trong khóa học nhập môn về khoa học dữ liệu tại Đại học Harvard.

Về cơ bản, luồng công việc của công việc Data Scientist thường là sự lặp đi lặp lại của năm giai đoạn:

### **Giai đoạn 1: Đặt ra những câu hỏi thú vị**

Trước khi bắt đầu quá trình nghiên cứu về dữ liệu khoa học, việc đầu tiên mà các Data Scientist cần làm là đặt ra những câu hỏi thú vị để xác định rõ vấn đề.

Điều này không hề đơn giản, vì sẽ có rất nhiều yếu tố, thông tin cần xem xét đến để đảm bảo giải quyết đúng mục tiêu khoa học của vấn đề. Một số câu hỏi mà các Data Scientist có thể đặt ra đó là:

* Mục tiêu khoa học của dự án này là gì?
* Ta sẽ làm gì tiếp theo khi có trong tay đủ mọi thông tin, dữ liệu cần thiết?
* Ta đang muốn dự đoán hay tính toán điều gì từ những dữ liệu khoa học ấy?

Việc trả lời các câu hỏi sẽ giúp Data Scientist hiểu hơn về mục tiêu cốt lõi của dự án. Hơn thế, những câu trả lời ấy còn đóng vai trò “xương sống” để xác định rõ những công việc tiếp theo.

### **Giai đoạn 2: Thu thập dữ liệu**

Lượng dữ liệu thu thập được sẽ đóng vai trò quyết định trong bất kỳ dự án khoa học dữ liệu nào. Chính vì thế, giai đoạn thu thập dữ liệu là công việc vô cùng quan trọng đối với bất kỳ Data Scientist nào.

Vì rất hiếm khi tất cả các dữ liệu ta cần được chắt lọc sẵn, thế nên ở giai đoạn này, các Data Scientist buộc phải thu thập càng nhiều mẫu dữ liệu càng tốt. Những dữ liệu nhiễu ấy sau đó sẽ được “làm sạch” để cải thiện chất lượng và giúp máy tính có thể hiểu và đọc được.

### **Giai đoạn 3: Khám phá dữ liệu**

Sau khi các dữ liệu được thu thập và có thể truy cập được, các Data Scientist cần dành nhiều thời gian để làm quen với dữ liệu, khám phá và thật sự thấu hiểu chúng.

Ở giai đoạn này, Data Scientist cần phát triển các giả thuyết về dữ liệu, đồng thời tìm kiếm các mẫu dữ liệu và “soi” những điểm bất thường. Bởi lẽ, dữ liệu không tự nói lên những thông tin chúng ta cần ngay lập tức, mà đòi hỏi người truy cập chúng phải phân tách, tổng hợp và phản biện với dữ liệu.

Mục đích chính của giai đoạn này chính là thấu hiểu dữ liệu, từ đó chuyển sang giai đoạn tiếp theo: Lập mô hình dữ liệu.

### **Giai đoạn 4: Lập mô hình dữ liệu**

Những dữ liệu được gạn lọc và “làm sạch” ở bước trên đôi khi không thể tự tiên đoán hay ước tính một xu hướng được. Chính vì thế, sau khi khám phá toàn diện dữ liệu, Data Scientist sẽ tiếp tục mô hình hóa chúng một cách rõ ràng, logic và dễ hiểu.

Việc mô hình hóa các dữ liệu sẽ giúp cả Data Scientist lẫn các bên liên quan có cái nhìn tổng quát về kết quả, tiên đoán hay “câu chuyện” đằng sau những dữ liệu đạt được.

### **Giai đoạn 5: Truyền đạt và hữu hình hóa kết quả**

Khi đã mô hình hóa tất cả những dữ liệu để phục vụ cho dự án khoa học, công việc tiếp theo của Data Scientist là truyền đạt và trình bày mô hình ấy một cách dễ hiểu.

Có thể nói, truyền đạt kết quả một cách rõ ràng, khúc chiết là một trong những kỹ năng quan trọng của một Data Scientist. Bằng không, các bên liên quan sẽ không hiểu được kết quả mà họ dày công thực hiện.

Và để truyền đạt rõ ràng nhất, một công việc khác trong giai đoạn này chính là hữu hình hóa (minh họa) kết quả. Bởi lẽ, các bên liên quan thường không phải là người hiểu rõ về khoa học dữ liệu, nên việc hữu hình hóa dữ liệu sẽ giúp họ dễ dàng hình dung hơn về kết quả, từ đó nhận ra những giá trị ẩn đằng sau kết quả của bạn.