Slide 22:

Cách xử lí dữ liệu nhiễu

adr: Vì chỉ có một giá trị âm nên sẽ hợp lý nếu thay thế nó bằng giá trị trung bình hoặc giá trị trung bình của cột adr, vì điều này sẽ không ảnh hưởng đáng kể đến việc phân phối dữ liệu.

người lớn: Khó có khả năng đặt chỗ với 0 người lớn, vì vậy 403 hàng này dường như có lỗi. Vì đây chỉ là một phần nhỏ của tập dữ liệu nên việc loại bỏ các hàng này là hợp lý.

trẻ em: Vì chỉ có một lượt đặt chỗ với 10 trẻ em nên nó có thể được coi là một trường hợp ngoại lệ và bị xóa khỏi tập dữ liệu.

em bé: Tương tự như cột trẻ em, vì chỉ có một lượt đặt chỗ với 10 em bé nên nó có thể được coi là một ngoại lệ và bị xóa khỏi tập dữ liệu.

slide 23:

Trong bài toán đặt phòng khách sạn này, biến mục tiêu là is\_canceled, cho biết liệu lượt đặt phòng có bị hủy (1) hay không (0). Cả kết quả dương tính giả (đặt phòng được dự đoán là đã hủy nhưng thực tế không phải vậy) và kết quả âm tính giả (đặt phòng được dự đoán là không bị hủy nhưng thực tế là như vậy) đều có thể gây tốn kém cho khách sạn.

Tuy nhiên, kết quả âm tính giả có thể tốn kém hơn vì khách sạn có thể mất khách hàng tiềm năng do đặt trước quá nhiều phòng và có thể sẽ không xảy ra trường hợp hủy phòng. Do đó, điều quan trọng là phải giảm thiểu các kết quả âm tính giả, có nghĩa là tối đa hóa việc thu hồi đối với lớp 'đã hủy' (1) sẽ là một cách tiếp cận tốt.

precision cũng rất quan trọng vì nó giảm thiểu kết quả dương tính giả, nghĩa là giảm thiểu trường hợp mô hình dự đoán việc hủy nhưng khách thực sự đã đến. Điều này có thể dẫn đến tình trạng đặt trước quá nhiều và khách hàng không hài lòng.

Do đó, điểm F1, là giá trị trung bình hài hòa của độ precision và recall, sẽ là một thước đo tốt để sử dụng vì nó cân bằng cả precision và recall. Cụ thể, điểm F1 cho lớp 'bị hủy' (1) sẽ là thước đo quan trọng nhất để đánh giá các mô hình trong dự án này.

Kết luận mô hình:

So với cả 5 mô hình, Randomforest cho thấy sự cải thiện về tất cả các chỉ số hiệu suất, đặc biệt là trong việc giảm số lượng Kết quả dương tính giả. Điều này cho thấy mô hình RF tốt hơn trong việc cân bằng sự cân bằng giữa precision và recall, dẫn đến điểm F1 cao hơn.

Nhìn chung, mô hình này không bị overfiting và đang hoạt động tốt trên tập test và dường như là một mô hình tốt để dự đoán số lần hủy, mặc dù vẫn còn chỗ cần cải thiện, đặc biệt là trong việc giảm số lượng Âm tính giả.

Như chúng tôi đã giải thích trước đó, số liệu quan trọng nhất cho dự án này là điểm F1 cho lớp '1' (đã hủy). Điểm F1 cao cho thấy sự cân bằng giữa việc giảm thiểu chi phí liên quan đến kết quả âm tính giả (đặt trước quá nhiều phòng dự kiến việc hủy sẽ không xảy ra) và kết quả dương tính giả (dự đoán số lượt hủy khi khách thực sự đến), điều này rất quan trọng để khách sạn tối ưu hóa lượng phòng trống trong khi giảm thiểu tình trạng đặt trước quá nhiều và khách hàng không hài lòng.