Heritage Health Prize Report

Tuan Nguyen

December 19, 2020

Chapter 1

Phân tích dữ liệu

Chương này mô tả về dữ liệu và một số nhận định trong quá trình phân tích, đồng thời dùng một số công cụ để thống kê và kiểm chứng nhận định.

1.1 Mô tả dữ liệu

1.2 Mô tả bài toán

1.3 Thống kê cơ bản

1.4 Khảo sát tương quan

Chapter 2

Data Preprocessing

2.1 Strategies

Dataset được cung cấp chứa dữ liệu claim của bệnh nhân trong 3 năm (Y1, Y2 và Y3). Mục tiêu của bài toán là dùng dữ liệu của 3 năm này để dự đoán số ngày nằm viện của bệnh nhân đó trong năm thứ 4 (Y4).

Tuy nhiên, chúng ta không thể có dữ liệu của Y4 để xây dựng mô hình từ 3 năm đầu tiên do chính sách của Kaggle. Vì vậy, trong giới hạn của project này, chúng ta sẽ dùng dữ liệu Y1 và Y2 để làm train set và dùng Y3 để làm test set.

Dữ liệu hiện có chứa thông tin các lần claim trong năm của một bệnh nhân. Để có thể dự đoán số ngày nằm viện cho một bệnh nhân trong năm kế tiếp, dữ liệu claim cần được xử lý phù hợp để chuyển tất cả thông tin claim thành thông tin tổng hợp của một bệnh nhân.

Khi quan sát tập dữ liệu, ta nhận thấy có một số dữ liệu cố định của bệnh nhân, không phụ thuộc vào năm, ví dụ: "Sex", "AgeAtFirstClaim". Các thông tin này không thay đổi dù bệnh nhân có claim bao nhiêu lần đi chăng nữa. Cùng với đó là một số dữ liệu thay đổi theo mỗi lần claim, chẳng hạn: "PayDelay", "PrimaryConditionGroup", ...

Do đó, sẽ có 2 chiến lược xây dựng dữ liệu cho mô hình:

Strategy A: Chỉ dùng dữ liệu của 1 năm để dự đoán cho năm sau. Các thông tin trong bảng Claims sẽ được tổng hợp theo từng năm.

Strategy B: Kết hợp dữ liệu của tất cả các năm trong quá khứ để dự đoán cho năm tiếp theo. Dữ liệu claim của Y1 và Y2 trong bảng Claims sẽ được kết hợp lại.

Trong project này, ta sẽ xây dựng cả 2 mô hình để đánh giá chiến lược nào sẽ cho kết quả tốt hơn.

2.2 Data Manipulation

Phần này sẽ mô tả cách thức xử lý từng feature để tạo ra dataset cho 2 chiến lược ở trên.

2.2.1 Bảng Members

Bảng này chứa dữ liệu chung của bệnh nhân và không phụ thuộc thời gian.

MemberID: Đã được de-identification, feature này chỉ được dùng làm index cho các feature khác và không dùng trong quá trình xây dựng mô hình.

AgeAtFirstClaim: Biểu thị giá trị rời rạc cho khoảng độ tuổi và sẽ được chuyển thành dữ liệu số là giá trị trung bình của khoảng đô tuổi đó. Xem bảng sau:

Giá trị ban đầu	Giá trị thay thế
0-9	5
10-19	15
20-29	25
30-39	35
40-49	45
50-59	55
60-69	65
70-79	75
80+	85
NaN	-1

Table 2.1: AgeAtFirstClaim

Sex: Giới tính của bệnh nhân sẽ được chuyển thành one-hot vector. Gồm 3 giá trị: M, F và NaN

2.2.2 Bảng Claims

Chứa dữ liệu claim của bệnh nhân, mỗi feature trong bảng này sẽ được xử lý theo Y1, Y2 và kết hợp giữa Y1 và Y2 (Y12).

ClaimCount: Đếm số lượng claim của mỗi bệnh nhân theo Y1, Y2 và Y12.

Provider: Đếm số lượng Provider duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

Vendor: Đếm số lượng Vendor duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

PCP: Đếm số lượng PCP duy nhất của từng bệnh nhân theo Y1, Y2 và Y12.

Speciaty: Đếm số lượng mỗi giá trị của Speciaty cho từng bệnh nhân theo Y1, Y2 và Y12.

PlaceSvc: Đếm số lượng mỗi giá trị của PlaceSvc cho từng bệnh nhân theo Y1, Y2 và Y12.

PayDelay: Dữ liệu sẽ được chuyển thành kiểu số. Giá trị 162+ chuyển thành 162. Sau đó tính toán các giá trị thống kê như min, max, avg, std, sum của từng bệnh nhân theo Y1, Y2 và Y12.

LengthOfStay: Dữ liệu được chuyển thành trung bình số ngày của mỗi giá trị và tính toán các giá trị thống kê như min, max, avg, std, sum của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

Giá trị ban đầu	Giá trị thay thế
1 day	1
2 days	2
3 days	3
4 days	4
5 days	5
6 days	6
1- 2 weeks	11
2- 4 weeks	21
4- 8 weeks	42
8- 12 weeks	84
12- 26 weeks	133
26+ weeks	182

Table 2.2: LenghOfStay

DSFS: Dữ liệu được chuyển thành số tháng và tính toán các giá trị thống kê min, max của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

CharlsonIndex: Dữ liệu được chuyển thành số và tính toán các giá trị thống kê min, max, avg của từng bệnh nhân theo Y1, Y2 và Y12. Xem bảng sau:

PrimaryConditionGroup: Đếm số lượng mỗi giá trị của PrimaryConditionGroup cho từng bệnh nhân theo Y1, Y2 và Y12.

ProcedureGroup: Đếm số lượng mỗi giá trị của ProcedureGroup cho từng bệnh nhân theo Y1, Y2 và Y12.

Giá trị ban đầu	Giá trị thay thế
0- 1 month	1
1- 2 months	2
2- 3 months	3
3- 4 months	4
4- 5 months	5
5- 6 months	6
6- 7 months	7
7- 8 months	8
8- 9 months	9
9-10 months	10
10-11 months	11
11-12 months	12

Table 2.3: DSFS

Giá trị ban đầu	Giá trị thay thế
0	0
1-2	2
3-4	4
5+	6

Table 2.4: DSFS

2.2.3 Bảng LabCount

LabCount: Chuyển giá trị 10+ thành 10 và tính toán các giá trị thống kê như min, max, avg, std, sum và số lượng claim (LabClaimCount) của từng bệnh nhân theo Y1, Y2 và Y12.

2.2.4 Bảng DrugCount

DrugCount: Chuyển giá trị 7+ thành 7 và tính toán các giá trị thống kê như min, max, avg, std, sum và số lượng claim (LabClaimCount) của từng bệnh nhân theo Y1, Y2 và Y12.

2.3 Dataset building

2.3.1 Strategy A

2.3.2 Strategy B

Chapter 3

Predictive Modelling

3.1 Explore Data Analysis

rgvrvrtvrgt

vrrgrtg

ddd **Nhận định:** dd

Chapter 4

Tuning Hyperparameter

4.1 Explore Data Analysis

rgvrvrtvrgt

vrrgrtg

ddd **Nhận định:** dd

Chapter 5

Conclusion

5.1 Giới thiệu

rgvrvrtvrgt

vrrgrtg

ddd **Nhận định:** dd

5.2 Kiến trúc hệ thống

5.3 Application Flows

5.4 Screenshoots